

In no one's best interest

The latest twist in Britain's ongoing scare about the safety of a widely used vaccine has brought researchers' conflicts of interest to public attention. Unfortunately, the affair has promoted a simplistic view of this complex issue.

Like it or not, we live in a world in which Mammon and science can walk hand-in-hand. Researchers often have a financial interest in the projects on which they work — they may own shares in a spin-off company, for instance, or hold a patent on a key discovery.

Dealing with these conflicts is one of the thorniest issues facing today's journal editors. Many journals, including the *Nature* journals, have taken the view that disclosure is the best policy — and ask their authors to complete conflict-of-interest statements.

Medical journals have blazed the trail in tackling this difficult issue. And at first glance, recent statements from *The Lancet* seem to continue this trend. When the journal was told that *The Sunday Times* newspaper had uncovered an undeclared conflict of interest relating to a paper it had published in 1998, its editors launched an immediate investigation. Two days later, they said that the paper should not have been published in its original form (see *Nature* 427, 765; 2004).

The paper in question described a novel syndrome combining gut inflammation and forms of autism (A. J. Wakefield *et al.* *Lancet* 351, 637–641; 1998). More controversially, it speculated that the combined measles, mumps and rubella (MMR) vaccine might be to blame — a link stressed in subsequent statements to the media by lead author Andrew Wakefield. This placed *The Lancet* at the centre of a storm. Worried parents have shunned the MMR vaccine. The medical establishment, citing a lack of evidence from epidemiological studies, has denounced Wakefield and the controversial paper.

The latest development is the allegation that Wakefield had accepted funding under Britain's legal-aid system to perform another study to investigate the possible link between the MMR vaccine and autism on behalf of parents who were considering suing for damages. Had they known this, say *The Lancet's* editors, they would have asked Wakefield to change the part of the paper that suggested there was a link.

Is this a reasonable response to the new allegation? Well, that's debatable. The link between MMR and autism was always speculative in the extreme, and many editors would have demanded its exclusion from the paper irrespective of any alleged conflict of interest. But *The Lancet* routinely publishes papers by employees of drug companies and by researchers who are paid to testify in legal cases relating to their work. Wakefield's speculation about MMR was not well grounded, and many experts in public health will be delighted that this prestigious journal is now distancing itself from this part of his paper. It's unclear, however, whether Wakefield was any more compromised by conflicting interests than are many other of *The Lancet's* authors.

Then there is the question of whether Wakefield really did hide his alleged conflict. He made no statement when submitting the original paper, but in correspondence published in *The Lancet* some two months after the original publication, he acknowledged evaluating a few children on behalf of the Legal Aid Board. Wakefield did not state the funding he received — now reported to be £55,000 (US\$103,000) — and there is some ambiguity as to whether he was referring to patients in the controversial paper. But his actions hardly seem those of someone determined to conceal a fatal conflict. And his belated disclosure did not prompt any public reaction from *The Lancet's* editors at the time.

The problem is that many observers of *The Lancet's* disavowal of Wakefield's speculation about MMR in the face of a journalist's inquiry will now assume that researchers' conflicts of interest inevitably undermine their integrity. But if these conflicts are managed properly, and disclosed, this needn't be the case. Few journals are in a position to aggressively police their conflict-of-interest policies, so further allegations are inevitable. It's to be hoped that the next editor to feel the heat does a better job of explaining the link between the alleged conflict and the validity of the paper concerned. ■

Ending the pain in Spain

Whoever wins the Spanish general election must deliver on their vague promises about supporting science.

In the run-up to Spain's general election, public funding of research has made a rare foray into the headlines. Politicians from across the spectrum have voiced their support for the country's scientists, following the release of a 'state pact for science' endorsed by the Spanish Society of Molecular Biology and Biochemistry.

The document, signed by 11 of Spain's leading scientists, calls for a doubling of the proportion of gross domestic product spent on research and development by 2010. It demands a similar expansion of the country's scientific workforce.

Both major parties say that they back these moves. But will the victors in the 14 March poll make science a lasting priority? They should have every incentive to do so. A large reservoir of talent exists — young Spanish researchers are highly welcome guests at labs throughout Europe and North America. But lack of political interest and Spanish bureaucracy have prevented most of them from returning to contribute to Spain's development as a scientific power.

Spanish politicians should wake up to the role science can play in

stimulating the economy. And they should look to emulate the few programmes and centres that point to a future in which Spanish science no longer punches below its weight.

Over the past three years, a programme named after Spain's first Nobel laureate, the neuroscientist Santiago Ramón y Cajal, has repatriated almost 2,000 of Spain's diaspora of postdocs. It should be expanded. But if this programme is to bear fruit, the returnees need to be given a working environment where openness, free movement of staff and stiff competition for funds are the norm. Unfortunately, this is not the case in Spanish academia.

In Madrid, however, the National Center for Cancer Research, or CNIO, a private foundation owned by the health ministry, has in just six years gained an international reputation that most of Spain's universities and institutes can only dream of. Bureaucracy is minimal, and scientists can be recruited at any time. In Barcelona, the regional Catalan government has set up institutes along similar lines. These successful experiments should be repeated on a much broader scale. ■





On the line
Marine biologists
add up factors that
damage fish stocks
p4



Ethical edge
Medical editors
contemplate a code
of conduct
p5



Under control
Biosafety protocol
tightens rules for
transgenic grain
p6



Sinking feeling
US budget cuts
threaten to pull the
plug on undersea lab
p8

Doctors battle to contain AIDS epidemic as unrest engulfs Haiti

Erika Check, Washington

Even as looters roamed the streets of the Haitian capital last weekend, AIDS doctors in Port-au-Prince were still seeing their patients and dispensing medicine. Support workers were visiting mothers with AIDS in small, central villages. And clinics in the rural centre of the country were giving out life-saving HIV treatments, thanks to an emergency shipment of a six-week supply of medicine.

Around Haiti, AIDS doctors say that their work is more urgent than ever before. About 6% of Haiti's adults are infected with HIV — the highest rate of infection in the Western Hemisphere. And some 30,000 Haitians die from the disease every year.

Determined researchers and physicians have established a network of clinics, community programmes and services to tackle the HIV problem at the grass roots. But in a country where the spread of AIDS has been driven by poverty and instability, the fresh political crisis threatens to fuel the epidemic. The turmoil is undermining programmes and studies that could yield crucial insight about what works best against the disease — the same doctors who run the treatment and prevention efforts also conduct vital research into the dynamics of Haiti's AIDS epidemic.

"The situation here is very, very difficult," says Raul Boyle, the Haiti coordinator for UNAIDS, the Joint United Nations Programme on AIDS/HIV. "Everything is totally paralysed. I think that normalization and the possibility for agencies to restart their operations depends on having some level of security ensured by the international troops."

Boyle and other doctors say that the current crisis could worsen the Haitian AIDS epidemic in several ways. Roadblocks have disrupted food and medical shipments around the country, and doctors are worried that interruptions to treatment will nurture biological resistance to AIDS therapies.

Anarchy and violence are likely to lead to higher rates of rape, and women in isolated areas who have no access to food for their families might be driven to form partnerships with soldiers or truckers. Such alliances have spread AIDS in other parts of the world, says Ruth Berggren, an AIDS specialist at



Out of control: chaos and uncertainty in Haiti threatens to exacerbate the nation's health problems.

Tulane University in New Orleans, Louisiana.

But physicians in Haiti are not giving up. Berggren works with MARCH, a group that uses mobile clinics to deliver food and medicine to 175,000 women around the central city of Mirebalais. It also runs a network of *accompagnateurs* — paid support workers who visit women with AIDS at their homes. Berggren says that roadblocks have forced MARCH to cancel some clinics. But she adds that *accompagnateurs* are still travelling on foot to check on their patients and to remind them to take their medicine. "This situation highlights how important the *accompagnateurs* system is," Berggren says.

In Port-au-Prince, doctors at the country's oldest AIDS hospital were forced to close their doors when looting and violence broke out in the city. But they switched to a contingency plan that allows them to continue consulting with patients by phone. The hospital, the Haitian Study Group on Kaposi's Sarcoma and Opportunistic Infections (GHESKIO), is one of the main recipients of Haiti's US\$67-million grant from the Global Fund to Fight AIDS, Tuberculosis and Malaria. The grant allowed GHESKIO to open 20 clinics around the country last year.

The money also pays for the distribution of anti-retroviral medicine to 1,800 patients.

Some of this reaches poor Haitians in the central part of the country through a network of clinics run by Partners in Health, based at Harvard University. Joia Mukherjee, medical director for the group, says that some of its clinics were closed two weeks ago when rebels seized control of towns where they were located. But he adds that doctors still saw 4,000 patients during that time.

Mukherjee says that poverty and political unrest are inextricably linked to the AIDS epidemic in Haiti and other countries hit hard by the disease. "If we give up on countries like Haiti, we're giving up on the epidemic," he says. "This kind of political unrest and structural violence is fomenting the epidemic worldwide."

GHESKIO's director, Jean William Pape, echoes Mukherjee's determination. He hopes that the US administration will follow through with plans to send more money to AIDS programmes in Haiti. US officials announced such plans on 23 February, just six days before President Jean-Bertrand Aristide fled the country.

"We have worked through the dictatorship of the Duvaliers, the turmoil of the transition period, the *coup d'état*, the embargo and continuous civil unrest. This is one more hurdle that we will need to overcome," Pape says. ■

Climate findings let fishermen off the hook

Quirin Schiermeier, London

Overfishing is not the sole cause of dramatically declining fish stocks in the north Atlantic Ocean, or worldwide, said marine biologists at a Royal Society meeting last week in London. Environmental changes such as climate warming may be just as important, they said, urging governments to consider these factors when managing fisheries.

"Marine ecosystems, particularly in the northern Atlantic, are much more vulnerable to natural fluctuations than previously realized," says Michael Heath, a biologist at the Scottish Fisheries Research Services' Marine Laboratory in Aberdeen, and UK chair of the international project Global Ocean Ecosystem Dynamics (GLOBEC).

The project has substantially improved understanding of the natural mechanisms driving marine productivity and population dynamics, Heath says. This must now be brought to the attention of decision-makers, he adds: "Our measure of success will be how much of the information gathered through GLOBEC is translated into policy-making."

Failure to reconcile ecology and commerce has been a hallmark of international fisheries policy for decades, said biologists at the meeting, which took place as the British government finalizes a widely anticipated report on the future of UK fisheries.

The report is expected to be published later this month by the prime minister's strategy unit and will set out Britain's fisheries management strategy for the next 15–20 years.

But marine biologists say that they have struggled to make their voices heard in the process. GLOBEC participants have provided



Net effect: fears remain over fisheries' futures.

some input to the report, which may criticize existing practices. But the fishing industry is lobbying equally hard to influence the report.

"The tensions run deep," says Dick James, chief executive of the Northern Ireland Fish Producers' Organisation. "Scientists tend to hold fishermen responsible for anything that happens, but we are certainly not to blame for environmental and climate change. Not all stocks show signs of depletion. We should fish what is there, and fish it responsibly, but we do not get credit for what is going up."

Scientists at the Royal Society meeting called for a concerted effort to ensure that fisheries remain sustainable and profitable.

"With or without environmental trends, fishing mortality, through commercial fishing pressure, is still too high and must be reduced," says Colin Bannister of the UK government's Centre for Environment, Fisheries and Aquaculture Science in Lowestoft.

Scientists involved in GLOBEC, the UK share of which is funded by £7 million (US\$13 million) from the Natural Environment Research Council, observed biological responses to environmental changes in marine ecosystems from the Baltic Sea to Antarctica.

"There is evidence for significant decadal-scale biological changes, which have major consequences for the abundance of natural resources," says Grégory Beaugrand, a marine biologist at the Sir Alister Hardy Foundation for Ocean Science in Plymouth.

Beaugrand and his team investigated the impact on the marine food web of varying water temperatures and wind strengths in the North Atlantic. They found that fluctuations in the abundance, size and composition of plankton result in long-term changes in the numbers of large, commercially important fish, such as North Sea cod (see *Nature* 426, 661–664; 2003).

In the past, herring populations in the Baltic Sea and cod stocks off Newfoundland have collapsed, and failed to recover, even after fisheries were closed — indicating that factors other than fishing were in play (see *Nature* 419, 662–665; 2002).

To develop a sustainable fisheries policy, it will be crucial to determine how much of changing mortality patterns is due to fishing operations, and how much to environmental trends, the meeting was told. "If we get this wrong, the penalty to society will be serious," says Martin Angel, a biodiversity researcher at the Southampton Oceanography Centre. ■

♦ www.pml.ac.uk/globec

♦ www.strategy.gov.uk/output/Page3854.asp

Bush sacks outspoken biologist from ethics council

Erika Check, Washington

An eminent cell biologist has been dropped from the US president's Council on Bioethics after publicly criticizing the administration's stance on stem-cell research.

Elizabeth Blackburn of the University of California, San Francisco, has been told that her two-year term will not be renewed. Three members, including a political scientist, have been added to the council.

Blackburn — discoverer of the enzyme telomerase, which plays an important role in ageing — had disagreed with the council's chairman Leon Kass on issues such as human embryonic stem-cell research, which she supports. She has also claimed that parts of the council's reports on stem-cell research and reproductive technologies have misrepresented scientific facts.

Blackburn says that the online journal

PLoS is about to publish a letter detailing her concerns. She thinks she was removed from the panel because of public disagreements with Kass and the US president George Bush. "I cannot think of any other possible reason," she says.

Critics of the Bush administration interpreted Blackburn's removal as the latest example of a tendency to fill advisory panels with members who support the administration's positions. On 18 February, the Union of Concerned Scientists alleged that such activities are undermining the administration's decision-making by distorting scientific information (see *Nature* 427, 663; 2004).

The Federation of American Societies for



Elizabeth Blackburn: thinks her critical approach led to removal.

Experimental Biology (FASEB), which represents US biomedical researchers, says that it is "extremely concerned" about Blackburn's removal.

"This decision undercuts the panel's capacity to make recommendations based on the highest quality of

information and a broad spectrum of viewpoints," FASEB president Robert Wells wrote in a letter sent to Bush on 1 March.

Kass, a bioethicist at the University of Chicago, says that Blackburn was not removed for her political views, but he declines to elaborate.

"We have people on the council who disagree with the president," Kass says. "He wants diverse opinions and he's got them." ■

Medical editors urged to accept ethical code

Jim Giles, London

Editors of medical journals should be as accountable for their actions as the authors they publish, according to a group of editors that has drawn up a code of conduct for scientific publishing.

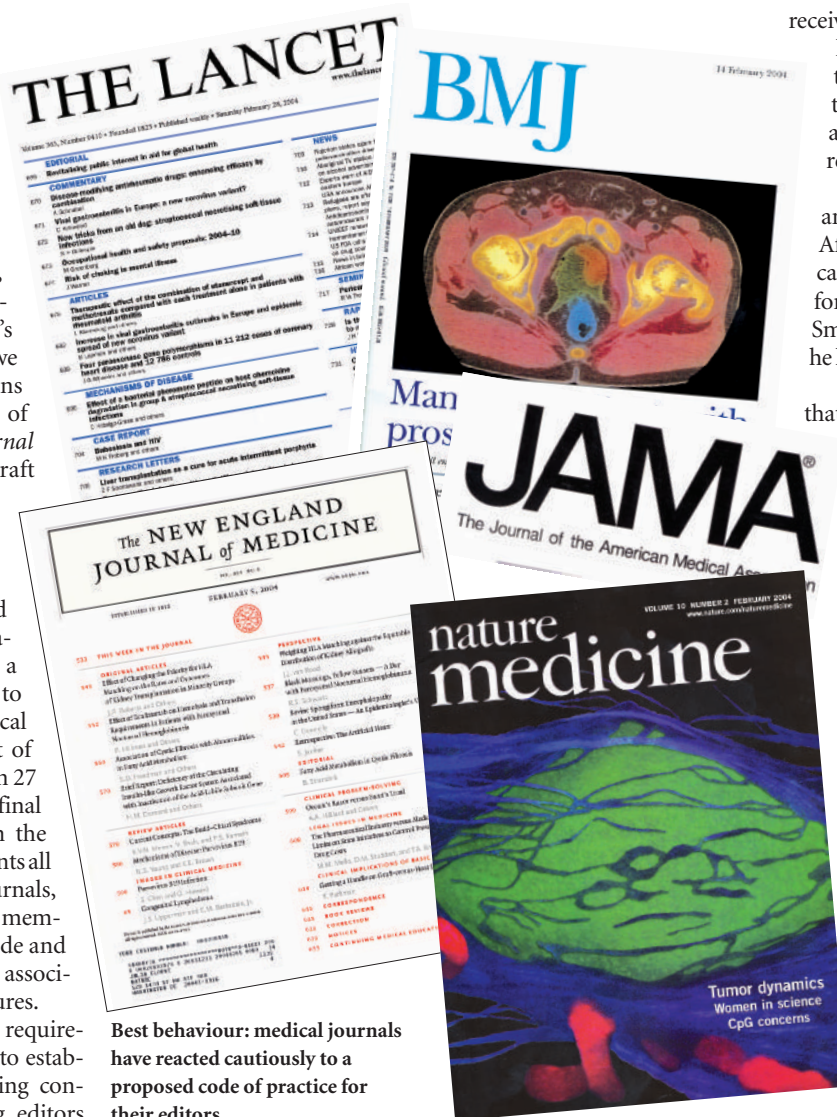
"Like everybody else, we are much more interested in other people's accountability than we are in our own," explains Richard Smith, editor of the *British Medical Journal* (BMJ), who helped to draft the new code. "Editors are perhaps some of the most unaccountable people in the world."

The London-based Committee on Publication Ethics (COPE), a body established in 1997 to help editors discuss ethical issues, published a draft of the code on its website on 27 February. It says that a final version will be ready in the next few months, and wants all editors of medical journals, including its 180 or so members, to sign up to the code and agree to be bound by the associated disciplinary procedures.

Many of the code's requirements, such as the need to establish systems for managing conflicts of interest among editors and other journal staff, are uncontroversial. Other parts call for more significant changes in editorial behaviour. When a journal suspects that the author of a paper has behaved unethically, for example, the code states that the editor must ensure that a relevant body, such as the author's employer, will carry out a proper investigation.

As COPE's annual report for 2003 reveals, editors face that situation quite frequently. One submitted paper, it says, revealed that to set up a control group for a study, blood samples were taken from healthy babies — a painful procedure that the paper's authors later admitted would never receive ethical approval. Another described how a patient with a serious disease was treated with a complementary therapy based on a plant extract, when a standard and effective conventional treatment was available.

Having received unsatisfactory explanations from the authors, editors from both journals concerned are contacting regulatory



Best behaviour: medical journals have reacted cautiously to a proposed code of practice for their editors.

authorities. "We used to just reject cases like this, but things have changed since COPE was established," says Smith, who is vice-chair of COPE's governing council. Smith believes that having access to papers that are not in the public domain allows journal editors to act as "privileged whistleblowers".

The code also confronts bad practice by editors themselves. COPE's 2003 report details several possible cases of this, including allegations that a journal made authors add references to other papers it had published in order to boost its impact factor. The editor approached COPE after being sacked for complaining to senior staff about the practice, and the committee is investigating the case.

Another allegation generated lengthy discussion when distributed on an e-mail list run by the World Association of Medical Editors. The case involved an essay that was accepted for publication in October 2001 by a medical journal. Four months later, the author

received a letter saying that a backlog of papers had forced the journal to reconsider the threshold for publication, and that the article had been rejected.

The story drew sympathy and anger from list members. After 25 posts, almost all critical, the offending editor came forward. "I cannot tell a lie," Smith wrote, as he admitted that he himself had made the change.

In his defence, Smith argues that his action best served the

interests of his readers.

The code does not cover the legitimacy of such decisions, but Smith points out that had it been in place, COPE would at least have had a framework within which to consider a complaint from the author.

Philip Campbell, editor-in-chief of the *Nature* titles, says he is currently considering the guidelines. But editors at some publications are uncertain about signing up. Drummond Rennie, deputy editor of the *Journal of the American Medical Association*, says that complaints about his publication are handled by a board of independent academics. He feels that this committee is already doing a good job of

dealing with complaints directed at editors. Rennie adds that the code would be more useful for smaller journals with fewer resources.

But members of COPE will have to go some way to compete with notable historical episodes of editorial wrongdoing. The committee's report cites the example of the late Cyril Burt, who founded the *British Journal of Statistical Psychology* in 1947 and was one of the pioneers of using twins to study the environmental and genetic origins of intelligence.

Some of Burt's studies have been discredited because of the apparent fabrication of data, and the COPE report points out that his editorial conduct, which included adding references to his own work to papers published by other authors, was not much better. "His *coup de grâce*," says the report, "came when he published a letter that he had written under a pseudonym, along with a response he also wrote himself under another pseudonym, so that he could attack a colleague."

♦ www.publicationethics.org.uk

US and biologists wary of strict biotech rules

David Cyranoski, Kuala Lumpur

How best to judge the potential environmental danger of a genetically modified organism? That's the question that dominated the first meeting of the parties of the Cartagena Protocol on Biosafety in Kuala Lumpur, Malaysia, last week. Attendees eventually agreed that detailed biological information should be provided with each shipment of genetically modified grain.

The outcome disappointed representatives of the largest exporters of genetically modified crops, notably the United States. Plant researchers were also worried that the proposed rules might reduce the sharing of materials.

It remains unclear whether tighter rules will slow the march of agricultural biotechnology, however — or simply render the protocol itself irrelevant. The United States and its allies are pursuing free trade in the technology through the World Trade Organization. If they succeed, observers say, it could render the biosafety protocol moot.

The protocol, provided for under the United Nations' 1992 Convention on Biological Diversity and enacted on 11 September 2003, aims to ensure the safe international shipment of what it calls living modified organisms. The meeting, which ran on 23–27 February, was the first attempt to decide how to implement this.

Delegates argued into the early hours of 27 February on how to enforce the protocol and how to manage compensation when damage occurs. The protocol is legally binding on the 87 countries that have ratified it.

The sharpest point of contention was the protocol's guidelines for labelling shipments of genetically modified organisms. The United States and its allies want to minimize additions to the existing rule for package markings to state that they "may contain"



Cold shoulder? US observers found little to smile about at the first biosafety summit.

such organisms. The United States, which isn't party to the protocol, but has been trying to observe it in practice, agreed with Canada and Mexico last October that exporters need not expand on this.

Anything more specific would hinder shipping, US representatives argued. "There are 12 or so varieties of genetically modified maize in commercial use — we have no idea what's in most shipments," says John Pitchford, who works in the US Department of Agriculture's office of international affairs.

But in Kuala Lumpur, representatives of the European Union and African nations successfully argued that exporters should put common and scientific names on each shipment, and information on the genetic

transformation event used to make each variety. Advocates say developing countries need this information for their own risk assessments. "We need to be able to check for ourselves," says David Hafashimana, a conservation biologist at Uganda's Ministry of Water, Lands and Environment.

Parties to the protocol will now begin implementing this approach, although the details won't be decided until another meeting in Germany later this year.

The few scientists among more than 1,000 negotiators and observers at the meeting expressed concern that the agreement will complicate the already fraught issue of sharing transgenic materials with colleagues abroad (see *Nature* 420, 602–604; 2002).

The agreement specifies that a significant amount of information, such as the transformation event and risk class, should be provided with all samples shipped for research. Hafashimana says that importing countries need this information: "In the case of an accident, how can we take appropriate measures if we don't know what it is?"

But Taesan Kim, a plant geneticist at the National Institute of Agricultural Biotechnology in Suwon, South Korea, is concerned that 'transformation event' could mean the number of insertions in a genome, the site of insertion, or even precise DNA sequences. "That's a lot of data," he points out.

And Kazuo Watanabe, a plant researcher at the University of Tsukuba near Tokyo, worries that cell cultures could attenuate while waiting in customs offices. "Researchers might become reluctant to send materials," he warns.

Gene-ecology agreement circles the globe

Thirty hours of flight time separate Tromsø in Norway and Christchurch in New Zealand. But researchers in the two cities have unearthed a common interest — gene flow — that springs from their respective positions as gateways to pristine polar regions. And last week, at the first meeting of parties to the biosafety protocol (see above), they agreed to team up to help other nations assess the risks of genetically modified organisms.

The Norwegian Institute of Gene Ecology (GenØk), based at the University of Tromsø, and the New Zealand Institute of Gene Ecology at the University of Canterbury in Christchurch, signed an agreement with the United Nations Environment Programme to help poor countries build the infrastructure needed to test genetically

modified organisms against environmental safety standards.

The two institutes have pioneered the new and contentious field of gene ecology, a discipline that includes the study of how consumption of transgenic foods affects the genes and long-term health of animals. "We start out by looking for differences where other groups assume everything will be the same," says Terje Traavik, scientific director of GenØk. The subdiscipline combines genetics, biochemistry, ecology and social analysis of related issues, he says.

The collaborators have received 5 million kroner (US\$700,000) for the project's first year from the Norwegian government, and hope this will be renewed annually.

David Cyranoski



Ready for action: Soyuz craft could maintain access to the International Space Station after 2010.

Ending of shuttle service puts space experiments 'at risk'

Declan Butler

Imagine your experiments are 365 kilometres above your head and there is no way to bring them back to Earth. That's the situation researchers using the International Space Station could face after 2010, when NASA stops flying the space shuttle.

The crew presents another problem. Scientists argue that the station needs six or seven people to perform a full research programme. At present, only three crew members can return safely to Earth on board a Russian Soyuz capsule in the case of an emergency — and in January NASA cancelled its Orbital Space Plane, which could have served as a lifeboat for a larger crew.

Officials from NASA and the other agencies that will operate the station set up a project team last month to explore options for maintaining access to it. The team is due to report in June.

"After 2010, the functionality of the space station risks being very limited," says John Kiss, a botanist at Miami University in Ohio and president-elect of the American Society for Gravitational and Space Biology. Most experiments must be brought back to Earth for analysis, he points out. "Astronauts are buckling experiments to their ankles," says Kiss, because room is already so restricted on return flights.

Transporting either cargo or crew to and from the station may become a huge problem, says Mary Musgrave, a plant physiologist at the University of Connecticut. She notes that NASA's fresh plans for manned missions to the Moon and Mars would use the station to test how organisms cope with

extended periods of low gravity. "That would require frequent and extended access to the station," she says. "At present we are not seeing a plan for providing that access."

"We have to find alternative ways to bring back scientific payloads," agrees Jorg Feustel-Buechl, head of manned flight at the European Space Agency (ESA). One possibility is to use ESA's Automatic Transfer Vehicle, designed to carry 7.5 tonnes of cargo to the shuttle and to burn up on re-entry. It could be modified to jettison a capsule that would parachute to Earth, but this would cost at least €100 million (US\$125 million), he says.

The crew problem is more acute. The last Soyuz capsule that Russia is contracted to send is due back in 2006, and two such capsules would be required to return a full crew. ESA officials are considering plans to buy more Soyuz craft and dock them two at a time throughout the station's lifetime. "We now have a clear rescue plan that permits a crew of six, which will allow the science to get done," claims Feustel-Buechl.

NASA has also pencilled Soyuz flights into its budget from next year until 2020. But there is an obstacle: a US law passed in 2000 bars the agency from buying additional Soyuz vehicles. Some Congress members are advocating an exemption to this rule that would let the agency buy Russian launches to keep the station going.

The project team's findings are expected to be considered when the heads of NASA, ESA and the other space-station partners meet in June to hammer out plans for the station in the light of NASA's revised exploration agenda. ■

Spanish lawmakers clash over control of stem-cell research

Laura Nelson, London

Spain's highest court is expected to rule this summer on a dispute between the national government in Madrid and the southern state of Andalusia over who should govern research on human embryonic stem cells.

The row echoes disagreements in the United States, where states such as California and New Jersey are trying to fund research that the federal government shuns.

The Spanish government passed a law in November that permits embryonic stem-cell research to be carried out under certain circumstances. Some researchers welcomed the law as relatively liberal for a predominantly Catholic country.

But Andalusia passed a still more liberal measure earlier in 2003. The province also backed ambitious plans to establish a stem-cell bank and research centre in Granada, with €5 million (US\$6.2 million) of regional-government money.

Both laws permit the use of stem cells derived from frozen embryos surplus to reproductive requirements. But the national law prohibits the creation of human embryos for research and states that embryonic stem cells should be used only as a last resort.

Researchers are worried that, under the national law, central government will have to assess and approve each project. "This might take six months or two years," says Francisco Martin, a stem-cell scientist at the Bioengineering Institute at Miguel Hernández University in Alicante.

"The national law was a reaction against the Andalusian law," says physiologist José López-Barneo of the University Hospital in Seville. "It was designed to stop the regional law."

Other researchers, however, say that the national law will clarify the situation countrywide, and is liberal enough to allow their work to proceed. They cite an agreement to collaborate on stem-cell research signed in January between the Spanish government and the Salk Institute for Biological Studies in La Jolla, California.

Juan Carlos Izpisua Belmonte, a geneticist at the Salk Institute, says that the agreement may permit researchers to do embryonic stem-cell research in Spain that would not get public funding in the United States. ■

Harvard seeks private \$100 million to build stem-cell centre

Washington Harvard University has announced plans for a multimillion-dollar centre for the study of human embryonic stem cells.

Rules established by the Bush administration in 2001 prevent federal dollars being spent on research into embryonic stem cells, so the university must raise private money to conduct the research. No official goal has been set, but researchers quoted in a 29 February *Boston Globe* article say that Harvard will try to raise approximately US\$100 million. Those funds will be used in part to hire up to 20 new faculty members for the facility.

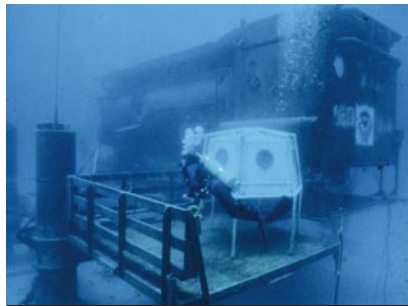
The centre is likely to study blood diseases, including AIDS, and conditions such as diabetes, neurodegenerative diseases such as Alzheimer's and Parkinson's, and cardiovascular disease. It will not just work on embryonic stem cells — studies on adult stem cells and animals are also planned.

Harvard's planned centre is the latest in a growing list of US research facilities that seek to bypass the federal ban on funding embryonic stem-cell research. Universities

in Wisconsin, Minnesota and California have already embarked on setting up privately funded efforts.

Budget cuts threaten to sink undersea lab

Washington Aquarius, the world's only habitable undersea laboratory, is facing possible closure following budget cuts aimed at the Undersea Research Program, a branch of the US National Oceanic and Atmospheric Administration (NOAA). According to programme director Barbara Moore, the government's proposed cut to US\$11 million — down \$1 million on last year's budget — is no longer enough to



Scuppered? Biologists studying Florida's corals may soon have to abandon the Aquarius lab.

guarantee that all of the lab's six undersea field centres can keep operating.

Aquarius, which is situated off Key Largo, Florida, is managed by the University of North Carolina at Wilmington, but gets about \$1.3 million from the government each year. In operation since 1988, Aquarius allows marine scientists to live 20 metres beneath the ocean surface, studying the coral reef for up to ten days at a time.

The Undersea Research Program has survived funding threats in the past — last year the House of Representatives sought to eliminate its budget, only to have the Senate restore it.

Galapagos fishermen call end to siege

London A week-long siege by fishermen of a research station in the Galapagos Islands ended last Friday after the Ecuadorian environment minister, Cesar Narvaez, signed a pact agreeing to review their demands. The fishermen had held 33 scientists captive in the Charles Darwin Research Centre, invaded National Park Headquarters, and blocked roads.

The government will set up a committee to review fishing laws enacted in 1998 to protect rare and endemic species such as turtles and sea-lions. The fishermen want to

end the ban on long-line fishing. But the demand is unreasonable and any review “will have to decide this is terribly destructive”, says Peter Kramer, president of the Charles Darwin Foundation.

Ecology centre aims to bridge Israeli–Arab divide

London A section of the fence separating Israel and Jordan has been dismantled to make way for a research institute. Construction work on the border-straddling Bridging the Rift Center, which will focus primarily on desert ecology, is due to begin on 9 March.

The centre has funding from US philanthropists and backing from Stanford and Cornell universities, which will train students and provide staff. The centre's first job is to compile a Library of the Desert, a catalogue of the species in its surroundings.

The institute's funds are channelled through the Bridging the Rift Foundation, established by Mati Kochavi, an investor with technology and real estate interests. Kochavi, born in Israel, says the foundation has raised tens of millions of dollars, enough to construct the first suite of labs, office space and accommodation and cover running costs for five years.



Water, water: Bangladeshis are threatened by sewage in surface waters and arsenic in wells.

UK geologists cleared over arsenic poisoning

New Delhi A geological survey team is not to blame for failing to detect arsenic in Bangladeshi wells, the Court of Appeal in London ruled last week. Bangladeshi villagers had accused the British Geological Survey of negligence.

In 1992, the survey team was called in to analyse Bangladeshi well water. Wells were dug in search of safer water. The country's surface waters are frequently contaminated with sewage and bacteria.

The study was not intended to look for arsenic, says the British Geological Survey. Only in 1995, when villagers began to show symptoms of poisoning, did experts realize that the water was polluted. “At the time of the study, there was no reason to expect arsenic to be there,” says David Lynn, director of science and innovation at the Natural Environment Research Council, which runs the British Geological Survey. Lawyers representing the villagers plan to appeal against the decision.

French gastronomy institute will teach the *art de vivre*

Paris The French government is to create an international school to award degrees in taste and gastronomy.

The Institute for Advanced Studies in Taste, Gastronomy and the Arts of the Table will open this autumn at the University of Reims, in the heart of the Champagne region. Researchers and lecturers will host sensory workshops and tastings, according to the prospectus. An online library will compile gastronomic works going back to the sixteenth century. The centre aims to promote the French *art de vivre*, and boost France's already strong food economy, according to the ministry of small and medium-sized enterprises.



A sleeping giant stirs

Mount Fuji is a cultural icon and Japan's most important geological feature. Yet until it began rumbling a few years ago, scientists had almost completely ignored it, says David Cyranoski. Is it preparing to erupt again?

Mount Fuji, Japan's tallest peak, last blew its top on 16 December 1707. It spewed for two weeks or more, raining down five centimetres of hot ash on the capital city Edo — now known as Tokyo — 100 kilometres to the east. In the countryside, ash filled the riverbeds, causing decades of frequent floods.

Today, such violence is rarely associated with Fuji. "Most people seem to think it is completely dead," says Toshitsugu Fujii, director of the Volcano Research Center at the University of Tokyo. A sacred mountain by tradition, it is the legendary home of numerous deities, and thousands of pilgrims still climb to the summit each year. Even the less religiously inclined compete for offices and apartments with a view of the mountain — visible from downtown Tokyo on clear winter days. "Everyone thinks it's pretty," says Masato Koyama, a geologist at Shizuoka University, about 40 kilometres southwest of the volcano. "They won't believe it can explode until it does."

But rumblings detected deep beneath Fuji four years ago reminded people that this giant is only asleep. Concerned that an eruption might be in the offing, the Japanese

government launched a concerted effort to study Fuji and learn when and how forcefully it could erupt again. Although the initial results seem to have produced more questions than answers, Earth scientists who once considered Fuji scientifically uninteresting now believe that it holds unexpected clues to the way volcanoes work.

A whole lot of shaking

In October 2000, Mount Fuji started experiencing waves of mild earthquakes more than 10 kilometres underground. They continued for about eight months. At their peak in April 2001, more than 100 earthquakes were recorded in one month, according to the National Research Institute for Earth Science and Disaster Prevention (NIED) in Tsukuba, vastly more than the average of 15 per year in the previous two decades. Although the quakes could not be felt at the surface, they carried a worrying message: the magma under Fuji is on the move.

If that movement portends an eruption, the consequences could be dire. "Nowhere in the world is a large capital city so close to a huge volcano," says Koyama. At the very least, dust and ash would bring Tokyo to a halt by



Digging into the past: with Mount Fuji (top) showing signs of waking up, scientists led by Toshitsugu Fujii (above) are rushing to gain a deeper understanding of the volcano by examining deposits from previous eruptions.

stopping traffic, grounding planes and even corrupting computer hard drives. But if the eruption came during summer, it could be much worse. Each year, 20 million people visit the golf resorts, amusement parks and summer homes that dot the volcano's flanks. Even more worrying is new evidence that Fuji has at times produced deadly pyroclastic flows — hot surges of gas, ash and rock that can travel at up to 150 kilometres per hour. Depending on the size of the eruption, these can cover distances from several kilometres up to more than a 100 kilometres.

In November 2001, the prospect of an eruption caused the Japanese government to act. It launched a ¥500-million (US\$4.5-million), three-year national project to develop a basic scientific understanding of Fuji. It also allocated a similar amount of money for the development of a 'hazard map' to detail



ROYALTY FREE/CORBIS

the likely threat to lives and property under various eruption scenarios.

Despite the volcano's fame, very little hard data on it existed, so scientists had to start virtually from scratch, as if the 3,776-metre peak had just been discovered. One reason for the dearth of information may be limited access to the mountain. The treacherous snowbound peak is off limits in winter, and the red-tape involved in getting samples out of the national park where Fuji is located is also forbidding.

But the other reason is scientific. "From a magma perspective I didn't find it very interesting," says Fujii. People who study magma tend to seek out purer sources, such as mid-ocean ridges, where the material comes straight from Earth's mantle, he explains. Fuji's magma, on the other hand, comes through a messy juncture where one tectonic plate bends under another, Fujii explains. "It's the dregs."

Core duties

Despite such reservations, Fujii leapt at the chance to lead the national Fuji research project, and within weeks he was drilling into the mountainside. The core samples taken from these 100- and 600-metre holes on the northeast slope are now being analysed. They should reveal in detail what sort of material erupted from the volcano in 1707, which in turn should help scientists to understand what caused that eruption. "To know what's going on today, we need to know what has happened in the past," Fujii says.

The evidence so far suggests that the 1707 eruption, known as Hoei for the period in which it occurred, does not fit neatly with standard models of volcanism. Records show that it was a particularly violent event. But Fuji consists almost entirely of a type of

volcanic rock known as basalt. Volcanoes of this type usually do not explode violently because, the theory goes, basalt flows freely and does not trap gas inside.

Fuji's explosive Hoei eruption seems to be an exception and so poses a challenge to the theory, says Yoshiaki Ida, a volcanic geophysicist at Himeji Institute of Technology in the western part of Honshu. "Perhaps the equation of low viscosity with non-explosiveness is mistaken," Ida suggests.

The Hoei eruption was strange in other ways, according to evidence from 300-year-old diaries. To most modern Japanese, the handwriting from that period is almost indecipherable, but Koyama, an amateur historian as well as a geologist, found historians who could make sense of the documents. He has pored over some 30 diaries that describe important features of the eruption, which lasted more than two weeks and was interrupted by a number of lulls. Mysteriously, the intensity of the eruption seems to have picked up in the last few days when it would have been expected to level off, says Koyama. "We need a mechanism concerning the magma that can explain this," he says.

Emergency planning

Meanwhile, the Sabo Technical Center, a semi-governmental institute based in Tokyo, is drawing up the hazard map. This will show in detail the most likely eruption scenarios: how much ash, lava and gas would be released; how long the eruption would last; and what the impact would be on the surrounding region. An accurate map depends in part on knowing what kind of emissions Fuji is likely to produce, so the Sabo team has been examining material from previous eruptions on Fuji's flanks.

What they have found has surprised geologists. Previously, evidence of 4,000-year-old pyroclastic flows had been found at two sites around Fuji, but those were thought to be a fluke of some kind. Basaltic volcanoes usually do not have pyroclastic flows. In 2001, however, Sabo surveyors dug up pyroclastic material at a different site. Initial estimates dated it to about 1,400 years old. "The find suggests that pyroclastic flows could be more common than previously thought," says map project leader Nobuo Anyoji.

The map-makers also need to know where the magma is, how it is moving and what forces act on it. But the nature of Fuji's magma chambers is proving elusive. In an effort to pin them down, Hideo Watanabe, Fujii's colleague at the Volcano Research Center, is leading the effort to monitor the volcano's seismic movements. Shock waves from quakes work almost like an underground sonar to produce an image of the size, position and movement of magma chambers. The greater the number of sen-

sors, the better that image will be. So Watanabe's team has nearly doubled the number of seismometers on the mountain's flanks to about 80, and placed three of them deep in the holes bored by Fujii to get a better signal-to-noise ratio than was previously available.

Once the arrangement of Fuji's magma chambers is understood, the deep low-frequency (DLF) earthquakes Fuji recently experienced may be easier to interpret. Scientists hope this understanding will help them to forecast and plan for eruptions.

Ready to rumble

Earth movements of various kinds have served as warning signs in the past. For instance, in 1991, shallow earthquakes and land deformation triggered a mass evacuation around Mount Pinatubo in the Philippines shortly before it erupted. Seismic activity also set off alarm bells before the 2000 eruption of Mount Usu on the Japanese island of Hokkaido.

But DLF earthquakes are much less reliable. For every example of DLF tremors preceding volcanic activity — several were detected before Pinatubo blew, for instance — there are many more where nothing happens. The result is a great deal of uncertainty over how to interpret Fuji's trembling.

"No convincing model has been proposed," says Motoo Ukawa, who heads the research project for predicting volcanic eruptions at the NIED. And the waveforms of Fuji's DLF quakes are particularly complicated, suggesting that the magma shifts causing them are complex and hard to model. The extra seismometers may help, but more time will be needed as the quake frequency has dropped back to previous low levels.

In the meantime, Fujii and his colleagues are getting ready to release their initial findings about the cores in May. Anyoji's group will also deliver the hazard map around the same time, but he says that it will be filled with more uncertainty than any he has ever worked on. "Where the eruption will take place, what kind of lava will flow, at what speed will it flow — these were all extremely difficult to predict," he says.

But if the magma that animates Fuji sounds like a force of evil, those who admire the mountain's aesthetics should give credit where credit is due. The 500 cubic kilometres of magma that Fuji has deposited on its flanks over the past 100,000 years has kept the mountain beautifully symmetrical in the face of constant erosion. Nevertheless, below the surface disaster lurks. "If the magma moves, sections would begin to collapse," says Richard Arculus, a geochemist at the Australian National University in Canberra. "It may be culturally sad, but that's the fate of all volcanoes."

David Cyranoski is *Nature's* Asian-Pacific correspondent.

► <http://hakone.eri.u-tokyo.ac.jp/vrc/VRC.html>

Just add water

Thanks to a sugar found in yeast, it may be possible to provide 'freeze-dried' blood cells to treat injured soldiers. The technique could also find applications in the cell-biology lab. Geoff Brumfiel reports.



The US military is one of the most bloodthirsty organizations on Earth. The troops hold regular blood drives to keep a required 70,000 units on hand at all times; and a veritable small army is needed to transport this blood to remote battle zones in Iraq or Afghanistan. It can take more than a week for refrigerated supplies to reach the field. That's a critical delay, explains Joe Bielitzki, a programme manager at the Defense Advanced Research Projects Agency (DARPA), which oversees speculative research for the Pentagon. "Typically, by that point nobody's bleeding," he says.

Ideally, the military needs blood supplies that can be stored and moved easily. So DARPA has assembled a team of US researchers to develop technology that will allow blood to be freeze-dried, rather like instant coffee, and stored at room temperature for years rather than days or weeks. It might seem an impossible task, but in just three years the group has achieved an impressive result — it has extended the shelf-life of human blood platelets, cells critical to wound healing, from a week to almost two years¹.

Medical applications aside, members of the DARPA team claim that their work could have wider uses in the laboratory. For example, it may be possible to store experimental cell lines for years at a time on a shelf, rather

than in expensive liquid-nitrogen freezers. And it could become easier to ship cells of all types, including precious embryonic stem cells, to and from labs around the world.

Cells and tissues can potentially be stored for long periods by freezing them, which slows their metabolism, and by dehydrating them, which removes the one ingredient essential to all biological processes — water. But more often than not, these techniques can also kill a substantial number of the cells. On a remote battlefield, freeze-dried packets of blood would be ideal — stable, lightweight and just requiring water to be reconstituted. But this means that the cells must survive the freezing and drying steps.

Kill or cure

Putting cells into a freezer exposes them to all sorts of danger. As water inside and outside the cells cools, ice crystals form and their jagged edges can rip the cells apart. Partial dehydration is a side effect of cooling, and if it is not controlled it causes the cells' membranes to shrivel and stick together. Rapidly cooling cells to liquid-nitrogen temperatures can prevent lethal ice crystals from forming by transforming the watery cytoplasm into an amorphous glass. But even if the cells survive freezing, the deathblow often occurs during thawing and rehydration,

when they are subjected to new stresses.

Improving these techniques is as much art as science, according to Juan de Pablo, a chemical engineer at the University of Wisconsin at Madison. Researchers typically treat their samples with chemicals, such as dimethyl sulphoxide, which help to stabilize the cell membrane and contents at low temperatures but have no protective effect during drying. The results are mixed, de Pablo says, and depend largely on the type of cell being preserved and the particulars of the technique, including cooling and warming rates. The chemicals, most of which are toxic, can trigger cell death themselves, and have to be washed away before the cells can be used.

Enter trehalose, a simple sugar found in organisms such as baker's yeast (*Saccharomyces cerevisiae*) and brine shrimps (*Artemia* species) that allows them to survive severe dehydration. Its properties are almost miraculous, says John Crowe, co-director of the Center for Biostabilization at the University of California, Davis, who has devoted most of his career to its study. "We've spent a lot of time looking at how it works," he says. It also has the virtue of being naturally non-toxic.

In the early 1980s, curiosity led Crowe and his team to begin probing the biophysical properties of trehalose. Then in 1994, the US Department of Defense offered funding



Sweet success: treated with a sugar called trehalose, some blood cells can be freeze-dried and kept for months or even years. This could make preservation of fresh blood (left) in liquid nitrogen (above) a thing of the past.



from disintegrating. Then, as the temperature falls below water's freezing point, trehalose forms an amorphous glass inside the cell, which prevents ice crystals from forming. If the cell is subsequently fully dehydrated, the glass becomes stable, and the cell can be kept at room temperature for long periods of time.

The main obstacle is that cell membranes are typically impermeable to sugars. "The big hurdle in the field is getting these molecules into the cells," says Mehmet Toner, a biomedical engineer at Harvard Medical School in Cambridge, Massachusetts. Toner's solution is to use pore-like proteins that dissolve in the cell membrane and can be opened and closed using chemical signals. Trehalose flows easily through these large pores, and Toner's group preserved 70% or more of human skin and connective-tissue cells using this technique². Another option is to genetically modify the cells to manufacture trehalose on their own³. But genetically modified cells have yet to be conclusively shown to work, according to Crowe.

Back to life

A simpler option is to warm the cells slightly to encourage endocytosis, a process that allows them to 'eat' the trehalose, before freezing the cells in a mixture of trehalose and proteins from blood plasma and drying them under vacuum. Using this technique, Crowe and Tablin have achieved 90% recovery of the platelets after rehydration — even after two years in storage¹. They have also scaled up the technique to produce whole bags of platelets at concentrations required for real transfusions.

Bielitzki is convinced that these techniques

will, within a matter of years, allow soldiers to carry their own supply of dried platelets into battle for use in an emergency. He also sees therapeutic uses for the cells for patients undergoing chemotherapy who have vastly reduced platelet counts. Being able to store bags of your own platelets in advance of treatment — and for longer periods — would ease transfusion problems later.

But using trehalose to preserve other cell types could prove much more difficult. Platelets are fairly simple cells that lack nuclei. Efforts to preserve more complex cells have so far had limited results. "There is still a way to go before we can desiccate these cells," says Toner. Key issues to be resolved are whether trehalose can get inside and stabilize the nucleus, and if it plays the same role inside the densely packed nucleus as it does in the cell's cytoplasm. Another important question is whether trehalose can prevent cell suicide occurring in more complex cells as a 'programmed' response to any mechanical or physical stress.

Crowe's group has so far been able to store red blood cells, which also lack nuclei, for about a week. "And even that is a big accomplishment," he adds. At the University of Wisconsin, de Pablo and his colleague Sean Palecek have been trying to store desiccated versions of the university's lines of embryonic stem cells, produced in the lab of James Thomson. "Stem cells are very finicky," says Palecek. "When they come under stress they usually kill themselves." To date, they have been able to preserve only about 10% of the stem cells they freeze. But the researchers believe that computer modelling might help them to understand how trehalose improves cell viability, and so refine their methods.

"Stem cells are very finicky — when they come under stress they usually kill themselves."

Human egg cells are also targets for the trehalose treatment. Eggs are much harder to freeze and recover than sperm or even embryos, so better survival rates would help women undergoing *in vitro* fertilization. In 2002,

Toner and his colleagues showed that using trehalose both inside and outside a frozen egg greatly increased the chances of recovery⁴.

Ultimately, Crowe and Tablin believe that trehalose may revolutionize the techniques used to store conventional cell lines in the lab. Cumbersome vats of liquid nitrogen would no longer be needed to preserve samples and, in many cases, it would be possible to send cell lines between labs with ease, says Tablin. "It sure would be nice to ship cells cross-country in an envelope," Bielitzki muses.

Geoff Brumfiel is Nature's Washington physical sciences correspondent.

1. Wolters, W. F., Walker, N. J., Tamari, Y., Tablin, F. & Crowe, J. H. *Cell Preservation Technol.* **1**, 175–188 (2003).
2. Eroglu, A. et al. *Nature Biotechnol.* **18**, 163–167 (2000).
3. Guo, N., Puhlev, I., Brown, D. R., Mansbridge, J. & Levine, F. *Nature Biotechnol.* **18**, 168–171 (2000).
4. Eroglu, A., Toner, M. & Toth, T. L. *Fertil. Steril.* **77**, 152–158 (2002).

to Crowe and another researcher at Davis, Fern Tablin, for new studies into trehalose and blood platelets.

"Platelets had never been available on the battlefield," says Tablin, who now co-directs the biostabilization centre with Crowe. Platelets cause blood to clot — essential for wound healing — but this means that they have to be stored separately from blood plasma at room temperature. Away from the battlefield, platelets normally have to be discarded after five days because of the risk of bacterial contamination.

After two decades of probing the structure of trehalose and how it interacts with cellular components, Crowe's team has worked out the main ways in which the sugar protects cells during drying and freezing. First, it replaces some of the water in the cell so that, as the temperature drops, trehalose prevents uncontrolled dehydration. Second, the sugar stabilizes the cell's membrane and stops it

Time to stop blaming communism in Hungary

Restrictive funding criteria put young researchers at a disadvantage in many countries.

Sir—Your News Feature “Dreaming on the Danube” (*Nature* 427, 94–95; 2004) is right to emphasize the unused intellectual potential in Hungary but it gives a flawed image of Hungarian science overall.

First, the tendency to equate Hungary with its capital Budapest is always annoying. Although Budapest is home to about 20% of the Hungarian population for historical reasons, most Hungarian students do not study in Budapest and most Hungarian researchers do not work there.

I am a 30-year-old chemist working at the University of Debrecen in eastern Hungary. In my discipline, the international reputations of the research communities at three Hungarian

universities (Debrecen, Szeged and Veszprém) are at least as good as, if not better than, that of their counterpart in Budapest. It is not accidental that the only Hungarian-born Nobel laureate who did his award-winning work in Hungary, Albert Szent-Györgyi, was a faculty member in Szeged.

Second, I think it is unhelpful to refer to the Hungarian language as “notoriously complex”. Its internal logic is certainly very different from that of European languages, but this does not make it more complex. Hungarian is a language in which there is only one past, one present and one future tense for verbs.

Finally, I disagree that the brain-drain problem has anything to do with Soviet-era thinking and the ‘old guard’. I agree

that the biggest obstacle to young Hungarian talent is the way in which research funding is distributed, but this problem is not unique to Hungary. Whenever grant recipients are primarily selected using criteria based on previous merit, such as the number of publications, it makes it difficult for young professionals to develop independent research.

Hungary has had two conservative and harshly anti-communist governments since 1990 and this system got their blessings as well. Blaming the old communist era for everything bad is a political game that has very little credibility now.

Gábor Lente

Department of Inorganic and Analytical Chemistry, University of Debrecen, 4010 Debrecen, POB 21, Hungary

Hungary: basic science needs European support

Sir—Your News Feature “Dreaming on the Danube” (*Nature* 427, 94–95; 2004) painted a grim picture of Hungarian science, reflecting the frustration of young scientists with the funding system.

Having returned to Hungary in 1995 and established a laboratory with mostly foreign support, I agree that without this it is almost impossible for young scientists to return and establish independent research. But the real problem is the serious underfunding of research, especially basic research, which affects scientists of all ages.

In my opinion, Hungarian science should be discussed in its context within the European Union (EU), which could play a much greater role in revitalizing science in a country with a record of producing high-impact research for low cost.

In 2003, the average Hungarian Scientific Research Fund (OTKA) grant was roughly US\$10,000, minus 20% for value-added tax. In my experience, it takes at least ten times more to establish a competitive life-sciences research laboratory in Hungary. Fortunately, many foreign grants offer funding in this range for scientists under 35 who want to return to Hungary, but Hungarian science cannot rely for ever on the generosity of foreign charities. The question is: in a small country with a low gross domestic product (GDP), where will the money come from?

When I returned to Hungary, I hoped that this help would come eventually from the EU. But the EU has indirectly made

matters worse, because the modest increase in Hungarian science funding has been used to emulate EU Framework programmes in supporting applied research, ignoring the needs of basic science and distributing disproportionately large funds to a lucky few. This approach is particularly problematic for a small country, which has its main strengths — as your “Eastern promise” Editorial points out — “not in technology development, where existing policy has its focus, but in basic research” (*Nature* 426, 369; 2003).

The same Editorial mentioned that the European Commission is finally starting to recognize the importance of the “untapped scientific potential” of newly joining countries for the competitiveness of the EU.

Currently, support for basic research is left to the member countries in the EU, not all of which can afford it. To allow fair competition for EU money in both applied and basic research, specific funds should be established to support basic science in new EU member countries whose per capita GDPs are well below the European average.

Private organizations in Europe and overseas already support basic science in Hungary. In my view, if the EU is serious about expanding the European research area, it must do more than simply drain talent away from the new member countries. The cost of such investment would be minimal, because 0.25% of the EU’s annual €178-billion (US\$227-billion) expenditure on research and development would create a fund ten times bigger than OTKA, and the EU could guarantee that the system would be fair and merit-based.

Then we could start working for Europe instead of dreaming about it.

László Hunyady

Department of Physiology, Semmelweis University, 1088 Budapest, Puskin u. 9, Hungary

Let’s hope there’s a good year on the cards

Sir—Yet another Christmas and New Year have gone by without a card from *The Journal of This* or *Trends in That*, in either mail box, real or electronic.

A little odd in a communications-conscious society, no? Perhaps I didn’t provide any ‘expert’ reviewing last year after all. Of course, I may just have unfashionable views of the festive season, coming as I do from the last century when people believed in Santa Claus and Mickey Mouse.

Come on, Editors, be a trifle old-fashioned for once, even if you do resort to e-mail.

Best wishes for what will surely be another glorious year in science.

Simon Wain-Hobson

Molecular Retrovirology Unit, Department of Virology, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France

Nature sent each of our reviewers for 2003 a copy of the last issue of the year with a letter of thanks — Editor, Correspondence.

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

Emerging epidemics

We have no one but ourselves to blame for the rise of new killer diseases.

Six Modern Plagues: And How We are Causing Them

by Mark Jerome Walters

Island Press: 2003. 206 pp. \$22

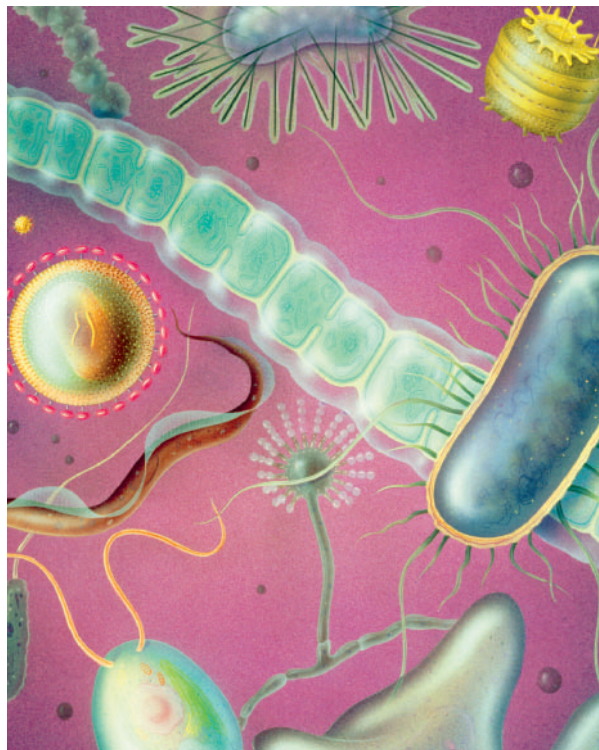
Tony McMichael

The word 'plague' has a profound historical, even biblical, resonance. In 1976, the historian William McNeill wrote his seminal book *Plagues and Peoples*, which explored how the emergence and spread of infectious diseases over millennia has reflected changes in human culture, demography and contact patterns. Since then, various other plague books — notably Laurie Garrett's *The Coming Plague* (Virago, 1995), Arno Karlen's *Plague's Progress* (Orion, 1995) and *Plagues* by Christopher Wills (Diane, 1996) — have appeared. Now we have *Six Modern Plagues* by Mark Walters.

Refreshingly, this latest book explores the underlying shifts in human ecology and behaviour that have potentiated recent epidemics. We are spared the usual combative clichés about conquering the microbes, being under siege and counter-attacking. Walters draws out of each example the fundamental lesson that, as we continue to change our ways of living, our environmental encroachments, our long-distance contacts and our food production methods, so we create opportunities for the genetically fleet-footed microbial world. Their drive to survive is as strong as ours, and is based on much longer experience.

Walters has selected six modern plagues from a list of several dozen infectious diseases that have appeared in human populations over the past three decades. As has been widely remarked, this is a very rapid rate of disease emergence. It reflects just how much modern humans have intensified our disturbance of, and interaction with, the natural world and its microbial multitudes. Walters quotes Peter Daszak, a leading conservation medicine researcher, as saying: "Show me almost any new infectious disease, and I'll show you an environmental change brought about by humans that either caused or exacerbated it."

Four of the featured infectious diseases — Lyme disease, Hantavirus pulmonary syndrome, West Nile fever and now mad cow disease (bovine spongiform encephalopathy, or BSE) — have recently emerged in the United States. The book also covers antibiotic-resistant *Salmonella* DT104, and HIV/AIDS. All six, says Walters, are "parables of the unintended consequences of careless human disruption of the natural systems that are our home". The emergence of *Salmonella*



Little cause for concern? Deadly diseases come in small packages.

DT104 and many other antibiotic-resistant bacteria has resulted from our injudicious use of antibiotics to enhance the force-fed growth of livestock — a practice that the US government, in particular, has been reluctant to curb because of vested interests and the commitment to market economics.

Mad cow disease emerged as a bizarre plague in British cattle in 1986. It subsequently transpired that the cause was bovine cannibalism, as rendered meat and bone meal from slaughtered sheep and cows were fed to cows as a protein supplement. The infective agent, a misshapen protein molecule (a prion) that imposes its peculiar molecular origami on other such proteins in the brain, causes neurological damage. It also affects human consumers of beef, causing variant Creutzfeldt-Jakob disease. The number of human cases, currently about 130, is several orders of magnitude fewer than the several million cows infected and culled.

Walters achieves a balance between environmental science, clinical medicine, human interest and social comment. The style is recognizably American, with each chapter presenting, in *Time* magazine fashion, a personalized illustration of a typical case of the disease in question. We meet, among others, a troubled UK dairy farmer; several young American deer-hunters who get a

relative of mad cow disease from eating affected venison (chronic wasting disease in deer); an HIV-infected chimp called Amandine; a Vermont cattle farmer at the point of death from *Salmonella* DT104; and a university student in New Jersey who caught Lyme disease from a tick lurking in the bushes around his frisbee arena.

Walters' book does not lead to specific prescriptive recommendations. Rather, he argues, insight into the underlying ecological processes that enhance microbial opportunism will enable us to tackle the roots of the problem, by "restoring the ecological wholeness upon which our health often depends".

In contrast to Walters' book is *The New Killer Diseases* by Elinor Levy and Mark Fischetti (Crown, 2003). Judging this book by its cover, it clearly is about Us and Them. The back cover states: "All around us — in our homes, workplaces, and public spaces — bacteria and viruses are evolving at a feverish rate, and our best defenses against them are in danger of being overwhelmed. The threat posed by emerging diseases is as formidable as any threat the human race has ever faced." The great challenge, it seems, is "to beat back the microbe armies".

Much of the book evinces this same idiom, telling us, for example, that "among the sinister threats we face in the pathogen war are deadly germs that infiltrate from abroad". And: "Just as with bacteria, killer viruses lurk everywhere, and they are even more cunning." This misconstrues the microbes' purpose. Microbes are interested only in enhancing their own survival and reproduction; they have no malign design on us. So, despite this book's lively accounts of recent outbreaks of infectious disease, suffused with human interest and lurid clinical detail, it fails to explain why infectious diseases are emerging or spreading at an increased rate in today's world. For that, see Walters' book. ■

Tony McMichael is at the National Centre for Epidemiology and Population Health, The Australian National University, Canberra, ACT 0200, Australia.

CARLYN IVERSON/SPL

Rooting out the wine plague

Phylloxera: How Wine was Saved for the World

by Christy Campbell
HarperCollins: 2004. 314 pp. £17.99

Jeffrey Granett

Grape phylloxera (*Daktulosphaira vitifoliae* Fitch) is an aphid-like insect native to North America that devastated European viticulture in the nineteenth century. On native American grapevines, the insect initiates galls on expanding vine leaves and immature roots, doing little harm to the plants' overall fitness. But when it finds itself on European vines (*Vitis vinifera*), its populations develop on mature storage roots, disrupting their function and introducing myriad secondary fungal pathogens, ultimately killing the vines.

When the insect first arrived in France in the mid-nineteenth century, the damage it caused could not be attributed to it because it was essentially invisible, as it lived and fed below ground. This was the last place a grape grower of the time would look for a cause of symptoms observed above ground. In addition, phylloxera are tiny (less than a millimetre in length) and populations vanish on dying vines.

Once the cause of the problem was discovered, there was no apparent solution. A large cash prize was offered for a cure and many off-the-wall ideas were tested, but the prize was never awarded. Removing and burning infested vines did not slow the spread. Carbon bisulphide fumigation of



Phylloxera ravaged Europe's vines until the use of American rootstocks led to French resistance.

infested vines helped, but was not economically viable in the long term. The vineyards were eventually saved by using phylloxera-resistant native American vines, not as direct fruit producers (the taste of wine from American grape species was unacceptable), but as rootstocks. The technology for this was optimized for vineyard locations and soil types, and has been amazingly successful ever since.

This story is dramatized in *Phylloxera: How Wine was Saved for the World* from the perspective of a wine lover. Christy Campbell focuses on how the people involved both in Europe and the United States went about solving the phylloxera problem. Science was clearly important, but so were the feelings and biases of the scientists and social progressives, the traditionalist wine growers,

those out to make money, and the wine lovers, including some who held positions of authority. In this context, the role of the scientists was not only to devise methods and test ideas, but to convince the rest of the stakeholders of the need for particular points of view and courses of action. Campbell tells how the key scientific players — Leó Lali-man, Jules-Émile Planchon, Louis Vialla and others — played the role of propagandists, sometimes successfully, sometimes not. Their's was a science that clearly required subjective as well as objective thought.

This approach to the phylloxera story is as important for the scientist to understand as it is for the layman. We must realize that even the most objective applied research is rarely convincing enough to find immediate utility without a lot of subjective interpretation, value judgements and soul searching. Campbell exemplifies this conflict with a lengthy description of the difficulties that the scientists faced in gaining acceptance of American vines as the solution to the phylloxera problem. American vines brought phylloxera to France in the first place, so how could their use be encouraged without increasing the risk of further phylloxera infestations and possibly opening up French viticulture to new, yet undiscovered scourges? How could the unacceptable 'foxy' taste of the American grapes not find its way into French wine, even if the American vines were used only as rootstocks? How can a noble French wine be truly noble and French if it owes a debt to wild, uncivilized American grapes?

The definitive answers to these questions had to wait until the rootstocks had been in the ground and producing wine for a decade or more. The first users of the rootstock technology did not have enough facts — they needed to be convinced by the propaganda. The story as told by Campbell has a colourful array of heroes and charlatans, and is neither completely linear nor holistic — but nor was the phylloxera problem caused or solved completely from the top down or bottom up. But it is an exciting, compelling tale. Campbell's short description of possible uses for genetic engineering in his concluding section make it clear that such stories are not over even after you put the book down. For English readers who would like more information, George Ordish's *The Great Wine Blight* provides more science, as do viticultural histories such as Harry Paul's *Science, Vine, and Wine in Modern France* (reviewed in *Nature* 387, 670–671; 1997).

Campbell's book is a fairly easy read, although the timeline gets a bit tangled occasionally. Its focus is on people solving a problem, but it provides enough science for the general reader to understand the controversies and the logic of the eventual outcome. ■
Jeffrey Granett is in the Entomology Department, University of California, Davis, California 95616, USA.

Reissued classics

The Sea Around Us

by Rachel Carson
Oxford University Press, \$45

Although Rachel Carson is best remembered as the author of *Silent Spring*, it was this natural history of the oceans that established her reputation as a science writer. The book was first published in 1951, so Carson was not able to include the subsequent discoveries of seafloor spreading and deep-ocean exploration. These events are included here in an introduction by Robert Ballard and an afterword by Brian Skinner.

America's Forgotten Pandemic: The Influenza of 1918

by Alfred W. Crosby
Cambridge University Press, £42.50 (hbk), £15.99 pbk

With bird flu and SARS regularly in the news, publication of this second edition of Alfred Crosby's account of the American pandemic

of Spanish Influenza is very timely. In 1918–19, Spanish flu killed more people across the world than the World War that preceded it. Since its first publication in 1976, this account of the devastation wrought by this disease has taken on more contemporary relevance, as the author reminds us in a new introduction.

The Piltdown Forgery

by J. S. Weiner
Oxford University Press, £8.99, \$14.95

The discovery in the early 1900s of Piltdown Man, the fossilized remains of an ape-like human, was greeted by much excitement. But a re-examination of the fossils in the 1950s revealed them to be forgeries. Joseph Weiner was one of three scientists who exposed this hoax, and his fascinating detective story, first published in 1995, remains the definitive account of the affair. As Chris Stringer points out in a new introduction and afterword, the identity of the hoaxer(s) is still not certain.

Science in culture

Science to a samba beat

Researchers and dancers joined hands in Rio in the name of Carnaval — and popularizing science.

Roald Hoffmann

For a week the city of Rio de Janeiro and its people put aside their class differences to craft an explosion of colour and music. A year of song-writing contests and work on costumes that cost months of wages lead up to informal block parades, parties, the incessant beat of samba, and a binge of popular culture. All in the middle of the Brazilian summer.

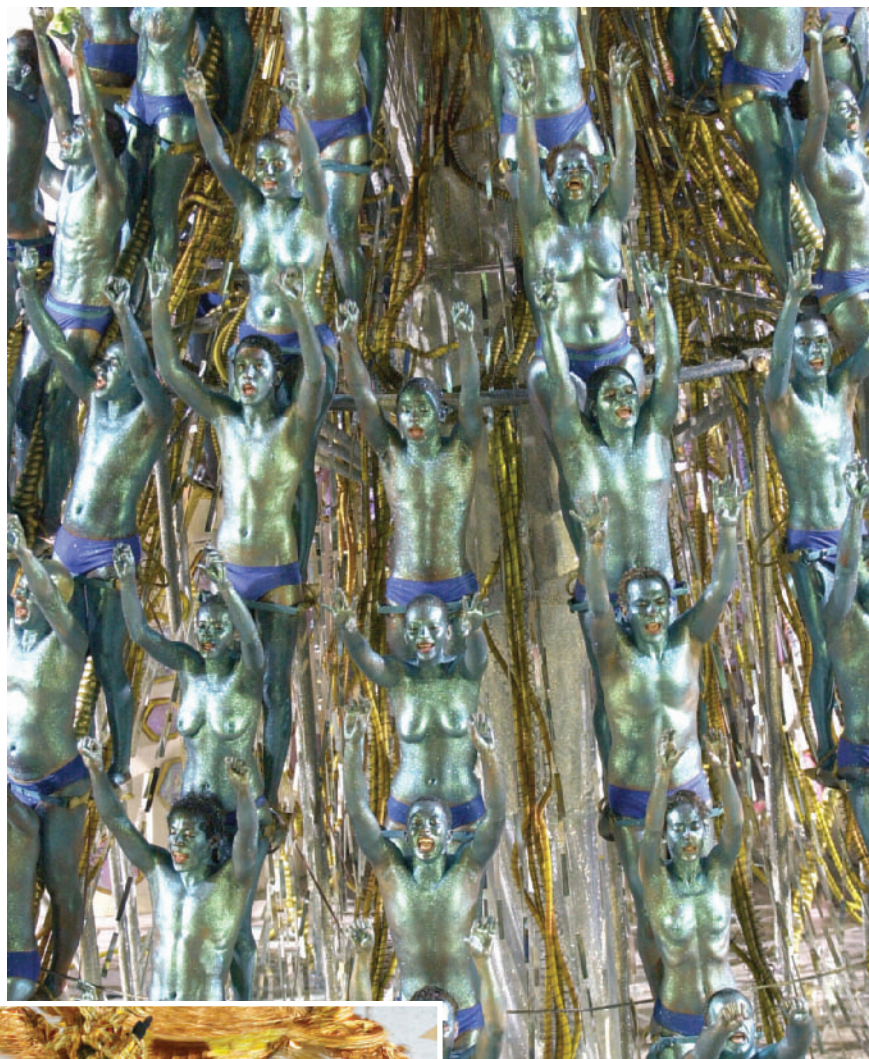
One of the high points of this festival is a competition between increasingly commercial samba 'schools', which parade down the Sambódromo, a structure like an elongated football stadium that seats 100,000. A billion more worldwide watch the performance on television.

The samba schools are judged by their theme, how well it is executed, and their spirit. And they are graded by their floats and their costumes (called *fantasias*), and on their song, the *samba enredo*, sung by a marching, walking and dancing throng of up to 5,000 people per school, and drummers playing the *bateria*. In a simpler time, some 50 years ago, Richard Feynman, dressed as a Greek, played a *frigideira* — a percussion instrument shaped like a frying pan — in a local parade.

Schools are upgraded or demoted on the basis of their ranking, and a neighbourhood's spirit depends on its team's placement. No wonder that the major samba schools hire a professional, known as a *carnavalesco*, to produce and direct their presentation. In Brazil this is considered a great profession.

For the first time, a major samba school, Unidos da Tijuca, chose a science theme for Carnaval — "The Dream of Creation and the Creation of the Dream: Art and Science in the Age of the Impossible". In elaborating on this theme, the *carnavalesco* and master alchemist Paulo Barros has worked closely over the past year with the team from the science centre at the Federal University of Rio de Janeiro, headed by Fátima Brito. This collective has much experience in popularizing science, but being asked to help in a Carnaval parade was something new. It's like asking scientists to advise on the half-time show at the Super Bowl — and the exposure of the human body on the floats in Rio is light years beyond Janet Jackson's flirtation with the risqué.

They took it on, the brave souls at the science centre, but not without trepidation. For there are doubters in Brazilian science who believe that the certifiably odd (I would say 'carnavalesque') representation of science in this Carnaval is but a distortion that adds to the public's misperception



Swayed by science? This striking float (above) — an allegory of human life — adds to the sound and colour of Rio's Carnaval.



of science. Some in the samba community also doubted that such a complicated theme would fly.

So what did the millions see on that hot night of 22 February in Rio? A fantastic float made of clock faces, driven by a popular actor, Carlos Palma, dressed as Einstein. Another striking float with 273 men and women in blue body paint choreographed in a representation of human life. People decked out as Dolly the sheep. And androids doing the samba down the *avenida*.

There was even an allegory of alchemy becom-

ing chemistry, including what looked like orbitals to this theoretical chemist — who, incidentally, was there in a Santos Dumont costume, balloons coming out of my back like angel wings, diplomatically dodging questions on who first discovered flight.

The process — a group of people intent on popularizing science in dialogue with a great samba school — was worth it by itself. But would Carnaval value this unique inclusion of science in popular culture? Three days after the parade, the judges ranked Unidos da Tijuca second (out of 14), its highest ranking ever. An analysis of the ratings by category reveals that the theme, and its ingenious, coherent execution, were responsible. I bet we see more science at Carnaval.

Roald Hoffmann is in the Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, USA.

The usefulness of parody

A timely rescue from a 'too clever' idea.

John Cairns

Cold Spring Harbor Laboratory has been the scene for magical turning points in the lives of many molecular biologists, and it quickly casts a spell on the temporary visitors teaching or attending the courses it runs each summer. For the permanent staff, all of whom could in those early days live on its idyllic Long Island campus, the effects are more complicated, simply because it is such a hard place to leave. My first impressions, as a temporary visitor, were negative. True, there was the luxury of being able to set up experiments while still in my pyjamas, return to our nearby apartment to read *The New York Times* in slippers ease, and then finally go back to the lab to harvest the yield from my now-ripe cultures. But this life seemed too pampered. Indeed, some years later my technician Paula de Lucia, after whom the gene *polA* was named, summed it up to some newspaper reporters in the scornful sentence, "The staff here do not even have to go shopping!"

In 1960, my family had come to Cold Spring Harbor so that I could spend a sabbatical year in Al Hershey's lab. He was the experimentalist who, more than any other and by his own solitary efforts, had generated the science of molecular biology. An austere figure, he made it his custom to work two days in every one — breakfast/work/dinner/sleep, and then a second round of the same starting late in the afternoon. He was a singularly silent man, and just about the only time you could get him to talk was when he was standing by the sink washing up glassware — apparently a task that he found was the only way to escape from the ceaseless burden of thought.

Faced with such a paragon, I felt I had to do something spectacular. Those were early days in the coming together of genetics and biochemistry. You could study either the genetics or the biochemistry of bacterial viruses, but only in a rather restricted sense could you directly link the behaviour of their genes to the behaviour of their DNA. I resolved to find a more direct way. Pleased with myself for having worked out a feasible (albeit complicated) method, I was explaining the whole thing to Bob Edgar, whom I happened to meet one evening as he was on his way to the weekly square dance that was held on one of the laboratory's lawns. Bob was one of Max Delbrück's cohort of young bacteriophage geneticists who was in Cold Spring Harbor to teach the Bacteriophage Course.



Picture perfect: the relaxed atmosphere of Cold Spring Harbor cuts through hierarchies.

Almost instantly, I came to regret having been so forthcoming. Edgar proceeded to convert my beautiful design into instructions for an imaginary square dance, prancing from one side of the road to the other as he acted out a dance of bacteriophage particles, each marked with a different isotope, pairing and separating, moving in lines, re-assorting themselves and so on. This lasted for the 300 yards between his classroom and the lawn with its waiting human dancers, where he plunged off into the golden twilight, leaving my experiment in tatters.

After some post-Edgar soul-searching I switched to a more straightforward project. Hershey (and Cyrus Levinthal at Columbia University) had just established that the DNA in each of the bacteriophage particles they were working with consisted of a single giant molecule. But it was not clear whether it was simply a double helix or something more complicated. Although the structure of DNA had been published several years before, the vision still seemed too good to be true. Was genetic information really stored in the linear sequence of the bases of DNA, and could it possibly be replicated in the way that Jim Watson and Francis Crick had suggested? Could the whole of genetics rest on something so simple? I had recently used tritium labelling as a way to localize the sites for the replication of

vaccinia virus DNA in mammalian cells. I resolved to use the same technique to measure the length of bacteriophage DNA, and thereby determine if it had the right length for a double helix or was shorter and therefore had a more complicated structure.

Thanks to Edgar's intervention, everything worked out well and the project nicely lasted out our year in Long Island. In the nick of time, he had made me see that the virtue of my initial project lay solely in its complexity, not in what it would actually tell anyone. Some years later, the pianist Glenn Gould delivered the same message at the end of his Fugue for Four Voices and Strings with the words "Never be clever for the sake of being clever!"

More importantly for my later stint as director of Cold Spring Harbor, Edgar's intervention made me see that one of the virtues of the lab — as someone once pointed out — was that its social structure is not 'vertically stratified'. I am not sure of the origin of this estimable trait, but I suspect it came from Max Delbrück. To give an example of his irreverence, I remember his saying how he was surprised to see the list of the new Nobel prizewinners and to realize that a whole year had passed since he had got his. "And I hadn't heard of any of them!" he added. ■

John Cairns is at the Clinical Trial Service Unit, Radcliffe Infirmary, Oxford OX2 6HE, UK.

Reconstructing the wild types

Edward F. DeLong

A challenging way to characterize the world's naturally occurring microbes is to piece together whole genomes from complex communities. An unusually acidic microbial habitat provides the setting for a ranging shot on that target.

Antonie van Leeuwenhoek discovered much of the microbial world by looking at it through a simple microscope. Later, while deciphering the roles of microbes in natural elemental cycles of sulphur, Sergei Winogradsky again showed the importance of direct observation of microbes in their natural habitats. These early lessons have not been lost on contemporary microbial biologists, either. For instance, over the past decade a variety of molecular methods have been developed that allow direct surveys and descriptions of naturally occurring microbial species in their native habitats¹.

Using the tools of production-scale genome sequencing and bioinformatics analyses, Tyson *et al.*² now take direct observation of native microbial assemblages to the next level. On page 37 of this issue, they describe the compilation of large portions of microbial genome sequences, retrieved directly from a natural community. The approach they used was whole-genome 'shotgun' sequencing. In this approach, total DNA from a complex microbial mixture is collected and smashed into tiny fragments, which are then pieced back together into their proper genomic arrangement with computer guidance. The results emphasize the potential of genomic strategies for studying complex microbial assemblages, as well as some of their inherent limitations.

To appreciate this feat, a little background on the environmental setting and the microbial players is useful. The microbial community studied by Tyson *et al.* is a relatively simple one, existing in an unusual, man-made habitat: the drainage conduits of an abandoned mine in northern California's Sierra Nevada mountains. This setting is classified as a 'superfund' clean-up site, as the resident microbes interact with pyrite (FeS₂) mine tailings to produce large amounts of sulphuric acid³. Here there are thriving microbial assemblages that have adapted to the self-generated low pH (about 0.5), moderate temperature (40°C) and abundant source of geochemical energy (FeS₂). The extremely acidic conditions and relatively restricted energy source combine to select for a relatively simple community, which Tyson *et al.* recognized as ideal for testing new genomic approaches in the environment. The main microbial players at the drainage site appear to make their living by



Figure 1 Microbial life in the pink. These apparently inert threads — shown close-up in the inset — are an example of the biofilm community that thrives underground, on the surface of drainage from the highly acid conditions in the mine sampled by Tyson *et al.*² (the Richmond mine at Iron Mountain, California). In this instance of 'environmental microbial genomics', Tyson *et al.* recovered two near-complete genomes of a bacterium and an archaean, as well as partial genomes of other microorganisms.

oxidizing iron and sulphide, and represent two of the three principal domains of life⁴, Bacteria and Archaea.

The genomic approach used by Tyson *et al.* consisted entirely of the whole-genome shotgun method⁵, as practised at large-scale DNA-sequencing centres such as the US energy department's Joint Genome Institute (JGI) in Walnut Creek, California, where the work was performed. After collection of an arbitrary chunk of microbial biofilm from the acid drainage site (Fig. 1), DNA was extracted, sheared to an average size of 3,000 bases, cloned and shotgun sequenced. Remarkably, the end-sequencing of a mere 50,000 or so clones (about 76 million base pairs) allowed the assembly of large, contiguous genomic regions from several dominant microbes in the mine-drainage community.

Following the DNA sequencing phase, the JGI group assembled a large fraction of the short DNA sequence reads into larger 'scaffolds'. The 1,183 scaffolds, totalling 10 million base pairs in length, could be sorted into distinct groups based on a combination of the read density (the number of individual

DNA reads per unit DNA length in a scaffold) and the percentage content of guanine and cytosine (G + C) bases. Sorting of the scaffolds using the combined criteria of G + C content and DNA read coverage grouped related genome sequence assemblies from the mix. Additionally, the location of a taxonomic molecular marker — ribosomal RNA (rRNA) — helped to link individual genomic scaffolds with specific microbial groups.

The high-G + C (55.8%) scaffolds having tenfold read coverage contained only one type of bacterial rRNA. This rRNA, which grouped with members of the genus *Leptospirillum*, encompassed a total of 2.3 million base pairs of DNA sequence. Similarly, the low-G + C scaffold group that had about tenfold read coverage contained one general archaeal rRNA type related to species of *Ferroplasma*, and totalled about 1.8 million base pairs. Their novel approaches to community assembly allowed the JGI bioinformatics group to stitch together large genomic scaffolds from these two abundant microbes in the mine drainage community.

T. JOHNSON

These preliminary results are truly encouraging. But conclusions about the origins, absolute genetic structure and identity of the DNA sequences must be considered as hypotheses that remain to be tested. The kind of data reported by Tyson and colleagues represents a sort of average genome structure, a patchwork quilt combining pieces of DNA from many individual cells in a complex population. For instance, an estimated 100 million *Leptospirillum*-like cells contributed to a total of 29,000 assembled DNA sequence reads for this group. Each DNA sequence that was read probably originated from a different (and potentially non-identical) cell within the population. So the assembled scaffolds are very different beasts from the singular genome sequences (determined from clonal laboratory isolates) that currently populate genome-sequence databases. Such environmentally derived genome-sequence assemblies need to be very carefully flagged up in the databases, so that their distinction as composite population assemblies will be obvious.

In the case of the *Leptospirillum* species, the nucleotide polymorphism rate (average sequence differences between individual reads) was low, about 0.08%. For this *Leptospirillum* population, then, there appeared to be only moderate genetic heterogeneity. In contrast, the *Ferroplasma*-like scaffolds showed a much higher polymorphism rate (around 2.2%) between individual reads that make up the scaffold. Given the locations and frequencies of these polymorphisms, the authors guess that the *Ferroplasma* polymorphisms arise predominantly from homologous recombination — a process whereby microbial cells can assimilate large blocks of foreign DNA into their genome.

There are, however, other plausible explanations for the patterns observed in the data. For example, the population may have originally been much more complex, but recently have undergone a 'selective sweep', leaving behind highly related but non-identical polymorphic genotypes⁶. What Tyson *et al.* presume to be genomic types that have been variously blended by lateral gene transfer may, in contrast, simply represent the visible survivors (by linear descent) of complex lineages — the surviving nodes in a family tree that has been extensively pruned by natural selection. At present, the existing data cannot fully distinguish between these alternatives. Better quantitative data on the levels of heterogeneity within and between *Ferroplasma* populations in this habitat, and a more quantitative assessment of genotypic representation in the environment, may help to clarify the issue.

But the complexities of deciphering genomic data from complex microbial populations are more of an opportunity than a difficulty⁷. After all, the reality is that

genomes are dynamic biological structures: any notion that they are singular, unchanging entities does not capture the process that shapes the diversity we see today. It is only by peering directly into naturally occurring genomic diversity, as Tyson *et al.*² have done, that the tempo, mode and mechanism of genome evolution and diversification, and its relationship to higher-order biological and ecological processes, will become clear. Our currently static snapshot views of microbial genomic diversity have the potential to develop into motion pictures as this emerging field develops. As predicted⁸, new vistas on the natural microbial world are

opening up dramatically, in part owing to the new capabilities afforded by modern genomic technologies. ■

Edward F. DeLong is at the Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, California 95039-0628, USA.
e-mail: delong@mbari.org

1. Pace, N. R. *Science* **276**, 734–740 (1997).
2. Tyson, G. W. *et al.* *Nature* **428**, 37–43 (2004).
3. Edwards, K. J., Bond, P. L., Gihring, T. M. & Banfield, J. F. *Science* **287**, 1796–1799 (2000).
4. Woese, C. R. *Microbiol. Rev.* **51**, 221–271 (1987).
5. Fleischmann, R. D. *et al.* *Science* **269**, 496–512 (1995).
6. Pals, T., Nakamura, L. K. & Cohan, F. M. *Int. J. Syst. Bacteriol.* **47**, 1145–1156 (1997).
7. DeLong, E. F. *Curr. Opin. Microbiol.* **5**, 520–524 (2002).
8. Woese, C. R. *Microbiol. Rev.* **58**, 1–9 (1994).

Superconductivity

Turn up the temperature

Piers Coleman

A more elaborate picture is developing of what makes some materials superconduct at relatively high temperatures. With it come hints for how to design materials with still higher transition temperatures.

Discovered a hundred years ago, superconductivity describes the flow of electric current without resistance in metals. Most metals do not become superconducting until cooled to within about ten degrees of absolute zero. But the

discovery, in 1987, of materials that become superconducting at much higher temperatures rekindled an old dream that one day room-temperature superconductivity might be achieved.

One of the mysteries of the high-

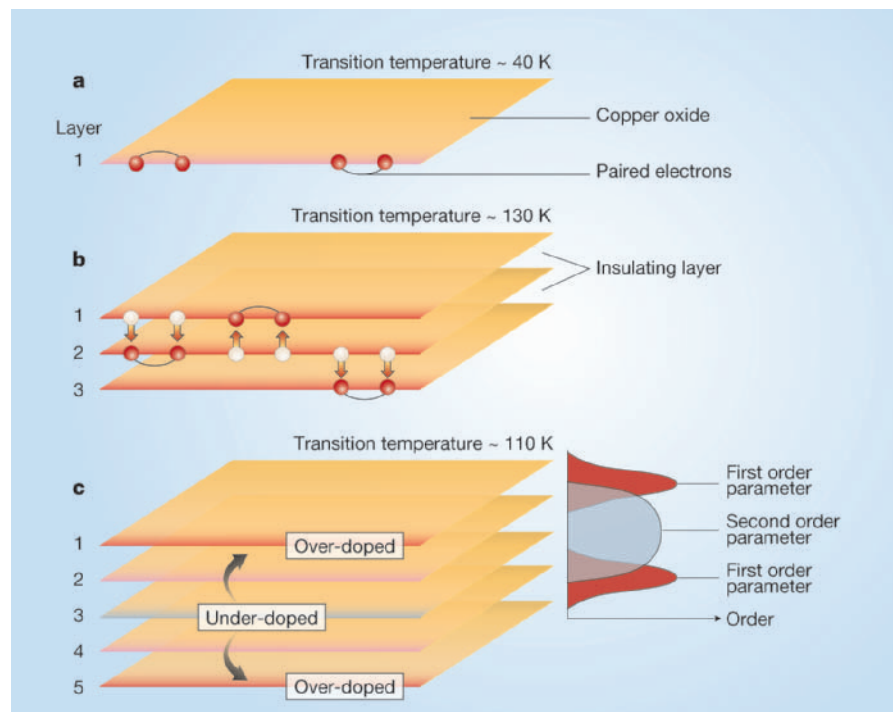


Figure 1 Layering effect. a, A single layer of lanthanum copper oxide becomes superconducting at a temperature of 40 K, as the electrons in the layer order into pairs. b, If the number of copper-oxide layers is increased to three (with layers of insulating material between them), the temperature at which the material becomes superconducting rises to 130 K, because electron pairs are now able to 'tunnel' between the layers. c, But if the number of layers is increased beyond three, the transition temperature does not continue to rise — in fact, it falls. Chakravarty *et al.*³ propose that the transfer of charge from the inner to the outer layers nucleates a second order parameter, in addition to the ordered pairing of electrons, and that this drives down the transition temperature.

Ecology

Chalk-hill blues

From the southern Mediterranean to the northern half of Europe, the distinctive colour of the chalk-hill blue butterfly *Polyommatus coridon* (pictured) is unmistakable — in contrast to the species itself, which has a rather patchy distribution. In the context of general concern over the effect that changes in climate and land use might have on biodiversity, T. Schmitt and G. M. Hewitt used this species, and another — *Erebia medusa*, the woodland ringlet — to test ideas about a species' history, its present-day patterns of genetic diversity and its robustness.

As they report in *Molecular Ecology* (13, 21–31; 2004), Schmitt and Hewitt have drawn on various datasets to conclude that the butterflies' location in Europe, determined by past climatic events, reflects their genetic diversity, and also, they speculate, their predisposition to population decline.



There are two major lineages of the chalk-hill blue in Europe, one in the east, the other in the west. After the last ice age, these groups probably emerged from two regions in the southern Mediterranean, and then underwent a northward postglacial colonization. Although the two lineages show equivalent genetic diversity, when the authors compared northern and southern populations of each, those further north were considered genetically poor, with respect to

diversity, and less stable demographically, than their southern counterparts. A similar pattern emerged for the woodland ringlet, only with a west–east axis.

So for both the woodland ringlets and the chalk-hill blues, the least stable populations — and so those with apparently less adaptability to change — are those furthest away from their former refugial areas. On the face of it, then, the butterflies' ability to adapt is a correlate of genetic diversity. If matters are to be taken further in examining the relationship between 'phylogeographical' patterns and population stability in European butterflies, more detailed genetic surveys will be required of these as well as other species — in addition to an assessment of the relative importance of genetic constitution and ecological change.

Katrin Bussell

in multilayer high-temperature materials the charge would redistribute, with the result that doping in the interior layers would in fact be less than in the outer layers. This idea has been confirmed by NMR measurements^{6,7} of the doping profile across the multilayer.

But central to the concepts underlying the new theory is the so-called order parameter. The Russian physicist Lev Landau was the first to realize that when matter develops new forms of order, the collective motion of its constituent particles can be described by variables that he called 'order parameters'. The advantage of Landau's order-parameter approach is that it means that the collective large-scale properties of a system can be easily separated from the gothic details of its microscopic motions. Using this concept, Landau and Vitaly Ginzburg⁸ were able to develop an order-parameter theory for the macroscopic properties of superconductors more than a decade before the pairing mechanism for superconductivity was devised by John Bardeen, Leon Cooper and Robert Schrieffer.

High-temperature superconductors are far more complex than their low-temperature counterparts, and there are many indications that their unique properties result from the competition between more than one type of order parameter. The electron correlations that are responsible for high-temperature superconductivity are still a mystery. But Chakravarty *et al.*³ bypass this unsolved problem by using the order-parameter approach to analyse the interaction between the superconductivity in each layer of the material.

The authors have melded these ideas of tunnelling, charge transfer and order into a simple Landau–Ginzburg expression for the total energy of the system. Looking beyond the order connected with correlated electrons, the basic idea of their theory is that, for more than three layers of copper oxide, a second, competing order parameter nucleates in the interior under-doped layers and lowers the superconducting transition temperature. By fitting the properties of a single-layer material and using the measured doping profile, they are then able to compute how the superconducting transition temperature evolves with the number of layers, obtaining good qualitative agreement with experimental data.

A key prediction of Chakravarty and colleagues' theory³ is that the 'pseudogap' should be larger in the inner layers of a multilayer superconductor. In normal superconductors, when electrons pair up, a gap develops in the electron energy spectrum; the higher the transition temperature of the metal, the larger the gap. But high-temperature superconductors break this rule: under-doped materials with lower transition temperatures actually develop a larger gap than optimally doped materials, even at high temperatures where they have not yet become superconducting. The nature of this so-called pseudogap is a

temperature superconductors is the link between their layered crystal structure and their transition temperatures (at which they become superconducting). These ceramic materials contain multiple layers of superconducting copper oxide, separated by insulating material — a kind of superconducting baklava. As physicists learned to synthesize more complex versions of these materials, with first one, then two, then three superconducting layers, the superconducting transition temperature shot up, from about 40 K in the single-layer materials to above 130 K in the three-layer materials^{1,2} (Fig. 1). But there it has stood for almost a decade — adding more layers unfortunately drives the transition temperature back down again. So what went wrong?

On page 53 of this issue, Chakravarty, Kee and Völker³ propose an explanation. Their theory brings together several insights into the physics of high-temperature superconductivity, and links three important effects — tunnelling, interlayer charge distribution and hidden order.

The first element of Chakravarty and colleagues' analysis³ is electron tunnelling between the superconducting layers. Inside a superconductor, electrons are bound into pairs. When two superconducting layers are brought close together, these pairs can

'tunnel' or jump from one to the other. (Here 'tunnelling' refers to the quantum-mechanical motion of a wave/particle through a region of space that would be forbidden in classical mechanics.) This tunnelling effect couples the two superconductors, and is called a 'Josephson coupling' after its discoverer.

About ten years ago (in a now disproven theory), Chakravarty, Sudbø, Anderson and Strong proposed that interlayer tunnelling is actually the origin of high-temperature superconductivity⁴. Chakravarty *et al.*³ have revived the idea of interlayer coupling, but now they argue that, although it is not the main engine of superconductivity in the layers, such coupling is responsible for the rise in transition temperature when going from one-layer to three-layer materials.

The authors³ also take account of how charge is distributed between the layers of the superconductor. A remarkable aspect of high-temperature superconductors is that, in many ways, these materials are closer to insulators than to metals. The mother compounds of high-temperature superconductors are insulating, and develop superconductivity only when extra charge carriers (electrons, or their positively charged counterparts, holes) are introduced into the material by chemical doping. Some years ago, Japanese physicists Ohta, Tohyama and Maekawa⁵ predicted that

matter of deep controversy⁹. But, according to one school of thought, it is produced by a second order parameter that competes with superconductivity, just as Chakravarty *et al.*³ envisage. There are various experiments under way¹⁰ that will shortly be able to test this idea and perhaps probe the nature of the hidden order inside the layers.

An attractive feature of this new theory³ is that it brings together several prevalent ideas about high-temperature superconductivity and provides a simple account of why the superconducting transition temperature first rises then falls as the number of superconducting layers increases. That we might now understand why the transition temperatures in these curious layered materials have been limited is exciting indeed — and from

this understanding, the prospect of designing materials with still higher transition temperatures looks more realistic.

Piers Coleman is at the Center for Materials Theory, Rutgers University, Piscataway, New Jersey 08854-8019, USA.
e-mail: coleman@physics.rutgers.edu

1. Ihara, H. *et al.* *Nature* **334**, 510–511 (1988).
2. Parkin, S. S. P. *et al.* *Phys. Rev. Lett.* **61**, 750–753 (1988).
3. Chakravarty, S., Kee, H.-Y. & Völker, K. *Nature* **428**, 53–55 (2004).
4. Chakravarty, S., Sudbo, A., Anderson, P. W. & Strong, S. *Science* **261**, 337–340 (1993).
5. Ohta, Y., Tohyama, T. & Maekawa, S. *Phys. Rev. B* **43**, 2968–2982 (1991).
6. Kotegawa, H. *et al.* *J. Phys. Chem. Solids* **62**, 171–175 (2001).
7. Kotegawa, H. *et al.* *Phys. Rev. B* **64**, 064515 (2001).
8. Ginzburg, V. L. & Landau, L. D. *Zh. Eksp. Teor. Fiz.* **20**, 1064 (1950).
9. Timusk, T. & Statt, B. *Rep. Prog. Phys.* **62**, 61–122 (1999).
10. Kotegawa, H. *et al.* *Phys. Rev. B* **69**, 014501 (2001).

Cell biology

A cellular choreographer

W. James Nelson

Specialized cells that form barriers such as those of the intestine adopt a distinct asymmetry. One particular protein may be a prime mover in bringing about such cellular organization.

The intestine is a hollow tube lined by finger-like projections (villi) which, in turn, are covered by a sheet of highly organized cells. As in other exposed surfaces such as the skin, these specialized epithelial cells are born at the bottom of the layer and move up to replace cells that are constantly being sloughed off the top. During this

time, the epithelial cells adhere tightly to each other within the sheet to provide a protective barrier between the ‘outside’ and ‘inside’ environments.

But the upper and lower surfaces of the sheet have different surroundings and carry out different functions. For example, the upper surface is involved in taking up

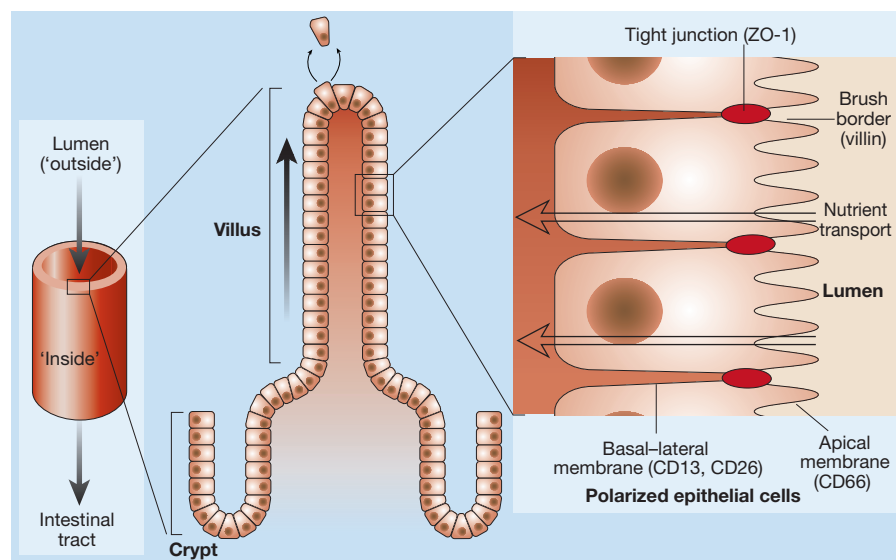


Figure 1 Structure of the intestine and epithelial polarity. The tube-like intestine is lined by thousands of villi. Each villus contains a crypt region in which new epithelial cells are born. These epithelial cells move upwards to replace cells that have been sloughed off. The epithelial cells in the villi are arranged in a highly ordered manner. On their apical side, which faces the lumen, is a brush border of microvilli, which contains villin, and apical proteins such as CD66 are present. Neighbouring cells adhere tightly to each other to form a barrier, through proteins such as ZO-1. Proteins such as CD13 and CD26 are present on the basal-lateral domain, underneath the lateral junctions between the cells.

nutrients from the food that passes through the intestine, whereas the lower surface transports these nutrients into the blood supply (Fig. 1). The corresponding cellular regions, the apical and basal-lateral domains, also contain different proteins. Knowledge of how this asymmetry, or polarization, is established is only just emerging. As they report in *Cell*, however, Baas *et al.*¹ have taken investigation further into what may be a major player in promoting cell organization in this context, a protein called LKB1.

LKB1 has already attracted our attention (reviewed in ref. 2). A tumour suppressor, its absence is responsible for Peutz–Jeghers syndrome, a rare human disease characterized by benign overgrowths of cells in the gastrointestinal tract³. Mice that express only one copy of *LKB1* also develop growths similar to those in human Peutz–Jeghers syndrome⁴. But why might loss of LKB1 function cause tumour growth? Normal cells have a limited lifespan, but cells lacking LKB1 grow continuously in culture until LKB1 is re-introduced⁴. LKB1 also regulates cell death (apoptosis), and growths from patients with Peutz–Jeghers syndrome have a low number of apoptotic cells⁵. LKB1 may be part of the well-characterized pathway of another tumour-suppressor gene, *p53*, which controls cell growth and apoptosis⁵.

Proteins that are similar to LKB1 are involved in tissue organization and polarization in a variety of lower animals^{6,7}, but Baas and colleagues¹ are the first to study what happens within the cells. This group previously identified STRAD α (STE20-related adaptor protein) as a protein that binds to LKB1 in the cytoplasm⁸. In their present study, using cells in culture that do not usually polarize, Baas *et al.* increased the production of STRAD α so that LKB1 accumulated in the cytoplasm (Fig. 2; LKB can be found in the nucleus and the cytoplasm²). Before LKB1 accumulated, actin filaments of the cellular skeleton criss-crossed the cell and several other proteins had random distributions. Afterwards, actin and another protein, villin, re-localized into a ‘cap’ on the apical cell surface that contained long projections of membrane (Fig. 2). These actin-rich structures looked superficially like the intestinal brush border, so-called because of the abundance of highly structured microvilli (Fig. 1).

The presence of a brush border is also a hallmark of epithelial polarity. Another protein, ZO-1, redistributed to a diffuse circle around the brush border when LKB1 was overexpressed. The distributions of several other proteins also changed when LKB1 accumulated in the previously unpolarized cells and resembled topologically those of polarized intestinal epithelial cells (Fig. 2). So several key characteristics of polarized cells were observed when LKB1 was introduced into cells that aren’t normally polarized.

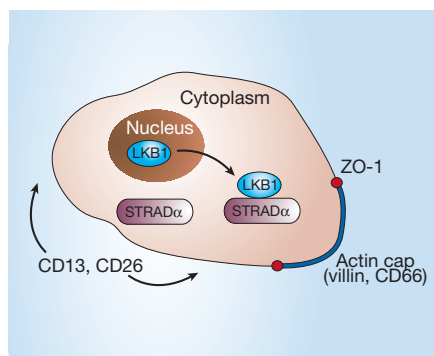


Figure 2 LKB1 and cell polarization. Some aspects of cell-surface organization that are evident in polarized intestinal epithelial cells can be induced by expressing LKB1 in non-polarized cells. LKB1 binds to STRAD α in the cytoplasm, and this is followed by re-localization of villin and the apical marker CD66 to an actin 'cap' with membrane projections. Around these projections is a ring of the protein ZO-1. CD13 and CD26 take up residence on the cell surface outside the actin cap.

Unexpectedly, this apparent ability of LKB1 to induce polarity occurred in single cells and in the absence of a solid support¹. Previous findings suggested that cells could not polarize until they had formed cell-cell junctions⁹ and that the surface¹⁰ to which they attached was also somehow involved.

Baas *et al.*¹ also examined the effects of depleting LKB1 in cells that form a fully polarized epithelium in culture. Given the dramatic effects of overexpressing LKB1 on protein distributions in the unpolarized cells, it might have been expected that LKB1 depletion would have a negative effect on cell polarity. But although the locations of some apical markers, such as actin and villin, were altered, they were not completely disrupted — perhaps these cells have additional, LKB1-independent mechanisms to maintain cell polarity.

Formation of the villin-containing actin cap is striking and may be central to LKB1-induced cell polarization. Perhaps activation of LKB1 by STRAD α (ref. 8) or PAR1 (ref. 7), another protein involved in cell polarity⁶, affects the behaviour of other proteins that induce the assembly of actin filaments and formation of membrane projections; villin would be in this pathway as its overexpression induces actin-containing membrane projections¹¹. These membrane projections may act as a platform for the assembly of other components necessary to generate a functional apical surface.

How are protein distributions altered when LKB1 accumulates? We don't know, but such redistribution might result from the directed delivery of proteins to different membrane areas; their retention in specific regions — for example, due to their ability to bind actin; or their exclusion from specific areas, such as the actin cap, and localization

by default to the rest of the membrane. This is particularly pertinent in light of the fact that Baas *et al.*¹ never found LKB1 to have a polarized distribution.

It remains unclear how LKB1-induced changes in cell organization relate to its roles in apoptosis and inhibiting cell growth. Baas *et al.* point out that changes in cell organization occurred before the effects on cell growth. So does LKB1 have several functions depending on whether it is localized in the nucleus or cytoplasm, or does binding to different proteins (for example, STRAD α or p53) or activation by proteins (such as PAR1 or protein kinase A¹²) direct LKB1 into different pathways?

W. James Nelson is in the Department of Molecular and Cellular Physiology, Beckman Center,

B121, 279 Campus Drive, Stanford University School of Medicine, Stanford, California 94305-5435, USA.
e-mail: wjnelson@stanford.edu

1. Baas, A. F. *et al.* *Cell* **116**, 457–466 (2004).
2. Boudeau, J., Sapkota, G. & Alessi, D. R. *FEBS Lett.* **546**, 159–165 (2003).
3. Jenne, D. E. *et al.* *Nature Genet.* **18**, 38–43 (1998).
4. Bardeesy, N. *et al.* *Nature* **419**, 162–167 (2002).
5. Karuman, P. *et al.* *Mol. Cell* **7**, 1307–1319 (2001).
6. Watts, J. L., Morton, D. G., Bestman, J. & Kempthorne, K. J. *Development* **127**, 1467–1475 (2000).
7. Martin, S. G. & St Johnston, D. *Nature* **421**, 379–384 (2003).
8. Baas, A. F. *et al.* *EMBO J.* **22**, 3062–3072 (2003).
9. Vega-Salas, D. E., Salas, P. J. & Rodriguez-Boulan, E. *J. Cell Biol.* **107**, 1717–1728 (1988).
10. Streuli, C. H., Bailey, N. & Bissell, M. J. *J. Cell Biol.* **115**, 1383–1395 (1991).
11. Friederich, E., Huet, C., Arpin, M. & Louvard, D. *Cell* **59**, 461–475 (1989).
12. Sapkota, G. P. *et al.* *J. Biol. Chem.* **276**, 19469–19482 (2001).

Molecular biology

Case of mistaken identity

Craig H. Bassing and Frederick W. Alt

An enzyme involved at grass-roots level in assembling genes for receptors that are essential in fighting infection has now been fingered as a suspect in certain cancers — proof that mistakes can be costly.

White blood cells called B and T lymphocytes play an important part in ensuring the integrity of the human immune system. The surfaces of these cells are home to a huge diversity of receptors, which recognize the foreign bodies that spark off our defences. The genes that generate these antigen receptors arise from a specialized form of DNA recombination known as *V(D)J* recombination. During this process, double-stranded DNA breaks are introduced at the variable (V), diversity (D) and joining (J) segments of the antigen-receptor gene loci by the RAG protein complex, and the cut segments are then joined together by general DNA repair proteins.

But RAG sometimes gets carried away, and may cut single-stranded DNA at other sites in the genome in a more haphazard way. In so doing, it could unwittingly create extra targets with which, if the breaks are not repaired, cut sections of the antigen-receptor loci might join. This, in turn, may be responsible for some forms of cancer of the lymphoid tissue, as parts of the antigen-receptor loci can fuse with, and activate, cancer-causing genes^{1,2}. On page 88 of this issue, Raghavan *et al.*³ report one such incidence of RAG's wayward behaviour. They find that single-stranded DNA breaks are generated in a particular genomic sequence near the *Bcl-2* proto-oncogene (the normal form of the gene that occurs before it becomes a cancer-causing oncogene). This causes the *Bcl-2* locus to fuse with a distinct part of the antigen-receptor locus. The

authors propose that this mechanism may underlie the development of human follicular B-cell lymphoma, a type of non-Hodgkin's lymphoma.

V, D and J segments are flanked by DNA sequences known as recombination signal sequences. These short sequences are specifically targeted by RAG — in particular, by its ability to generate double-stranded breaks⁴. Normally, *V(D)J* recombination only occurs between gene segments that are flanked by complementary recombination signal sequences, which are referred to as 12-RSS and 23-RSS. In the first phase of the reaction, RAG induces a single-stranded nick between two of the participating *V, D* or *J* segments and their flanking recombination signal sequences. It then converts the initial nick into a double-stranded break. In the second part of the reaction, so-called 'non-homologous end-joining' proteins process and join the RAG-held coding end and the recombination signal sequence ends to form coding and recombination signal sequence joins. The fused recombination signal sequences are considered a signature of bona fide *V(D)J* recombination (Fig. 1a, overleaf).

In some human lymphomas, translocations (which involve the aberrant movement and fusion of regions of DNA) can occur. They can arise when *V(D)J* recombination takes place between chromosomes targeted by recombination signal sequences in antigen-receptor loci and fortuitously complementary (cryptic) recombination signal sequences near a proto-oncogene. The best

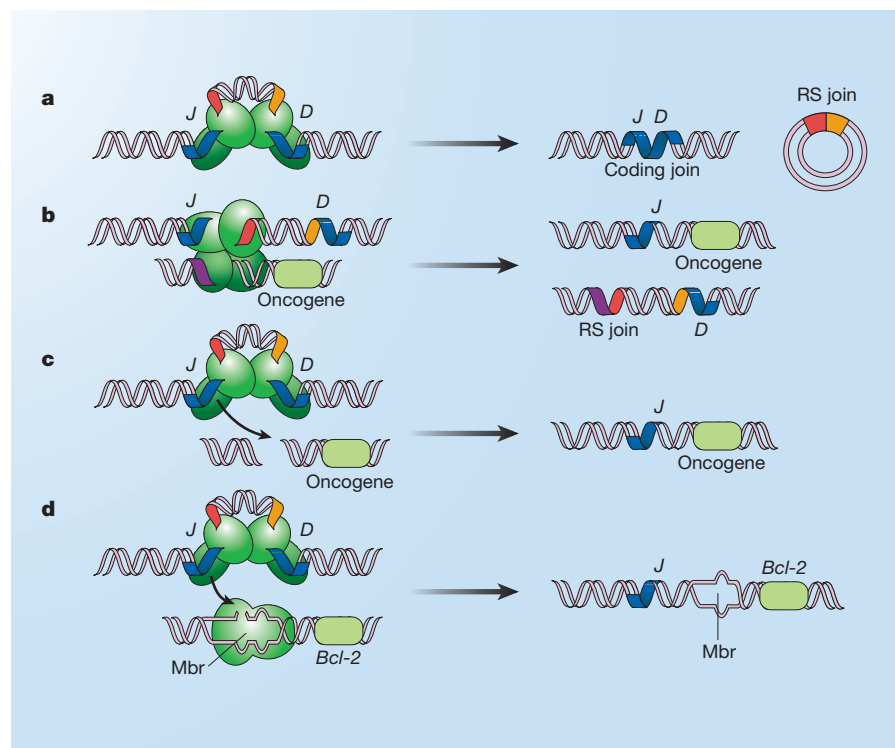


Figure 1 Aberrant $V(D)J$ recombination and translocation. **a**, Recombination between the variable (V), diversity (D) and joining (J) segments of antigen-receptor genes is catalysed by the RAG protein complex (green; only D and J are shown here, blue). RAG targets complementary recombination signal sequences, 12-RSS (orange) and 23-RSS (red). These sequences and the intervening DNA are cut out and join to form a DNA circle. The remaining coding sequences are also joined. **b**, RAG-mediated translocations between chromosomes. RAG can mediate translocations by catalysing $V(D)J$ recombination involving the recombination signal sequence of a bona fide antigen-receptor locus and a complementary, but cryptic, recombination signal sequence adjacent to an oncogene. Such translocations can occur through several related pathways. **c**, Coding or recombination signal sequence ends may escape from chromosomal RAG post-cleavage complexes and fuse with double-stranded breaks at other locations (translocation targets), sometimes near oncogenes. **d**, Coding or recombination signal sequence ends may escape from chromosomal RAG post-cleavage complexes and fuse with RAG-initiated lesions induced at the *Bcl-2* major breakpoint region (Mbr). The RAG-mediated nicks in the Mbr are hypothesized to arise as a result of the unusual DNA structures.

evidence for this is the finding of fused recombination signal sequences on reciprocal translocations^{5,6} (Fig. 1b). However, cryptic recombination signal sequences are not always present, leading to speculation that translocations can result when RAG-generated double-stranded breaks at antigen receptor loci on one chromosome join aberrantly with non-RAG-generated breaks on the other chromosome through some DNA end-joining pathway^{1,2} (Fig. 1c).

Various factors have been proposed to cause such putative double-stranded breaks to become a target for translocation; these include replication errors, sites that are particularly fragile, stress caused by certain species of oxygen, and short-wavelength, high-energy radiation². But some translocation breakpoints in antigen-receptor loci cluster within relatively short regions⁷ — the aptly named major breakpoint region (Mbr) of the *Bcl-2* gene is one such example. So it's quite feasible that some DNA sequences might suffer a higher level of random breakage than others, possibly

because they are targets for DNA-cleaving enzymes called nucleases⁸.

Raghavan *et al.*³ tested the notion that RAG itself may misrecognize and cleave the *Bcl-2* Mbr. They incorporated Mbr sequences into a substrate for $V(D)J$ recombination and investigated whether the sequences could function as a 12-RSS or 23-RSS. Consistent with an earlier study⁹, the authors found that Mbr sequences do not function as a recombination signal sequence. However, recombination substrates containing both the Mbr and also a set of paired recombination signal sequences accumulated a significant number of RAG-dependent breakpoints between the Mbr sequence and the more distal recombination signal sequence. This suggests that the translocations arose because the normal $V(D)J$ recombination reaction was misdirected to the Mbr. Purified RAG proteins could even specifically nick Mbr sequences in DNA vectors, but only on one strand. In the latter context, Raghavan *et al.*³ found that chromosomal Mbr elements often adopt an

unusual single-stranded DNA structure. Based on these findings, the authors suggest that coding ends released from a normal RAG cleavage complex at the locus encoding a certain part of the antigen receptor — the heavy chain — might recombine with a RAG-induced *Bcl-2* Mbr lesion (Fig. 1d). Raghavan *et al.* also propose that RAG-mediated cleavage of stable, unusual DNA structures in the *Bcl-2* Mbr could provide the molecular basis for the translocation event — between chromosomes 14 and 18 — that occurs in follicular lymphomas.

Although these findings are intriguing, most of the data come from *in vitro* modelling and cell culture. It has thus not been determined unequivocally that RAG is the nuclease that confers genomic instability on the Mbr in developing B lymphocytes. As the authors note, some unknown nuclease, perhaps attracted by the unusual DNA structure of the Mbr, might induce *Bcl-2* Mbr lesions. In any case, providing unequivocal experimental proof that RAGs carry out this aberrant function in the context of the *Bcl-2* Mbr translocations *in vivo* may prove a very difficult, if not impossible, task.

But even with this lingering question, Raghavan and colleagues' findings³ have important implications. The human genome is littered with cryptic recombination signal sequences and with sequences that, like the *Bcl-2* Mbr, may form unusual structures that are misrecognized and nicked by the RAG proteins. So RAG-initiated single-stranded nicks could theoretically contribute more broadly to the generation of genomic instability and tumour formation during lymphocyte development.

Finally, the process of class switch recombination that occurs in mature B lymphocytes, whereby the class but not the specificity of the antibody produced by the lymphocyte is changed, also involves the generation of single-stranded DNA structures. These structures then become substrates for the lesion-inducing activities of activation-induced deaminase¹⁰, the nuclease responsible for class switch recombination. Thus, the occurrence elsewhere in the genome of sequences predisposed to the generation of single-stranded DNA might also fuel fusion events with structures produced at the heavy-chain locus by activation-induced deaminase. Such processes would be analogous to those proposed by Raghavan *et al.*³ to mediate aberrant RAG translocations to the *Bcl-2* Mbr. ■

Craig H. Bassing and Frederick W. Alt are at the Howard Hughes Medical Institute, The Children's Hospital and Department of Genetics, Harvard Medical School, and the CBR Institute for Biomedical Research, 300 Longwood Avenue, Boston, Massachusetts 02115, USA.
e-mails: alt@enders.tch.harvard.edu
bassing@cbr.med.harvard.edu

1. Roth, D. B. *Nature Rev. Immunol.* 3, 656–666 (2003).

2. Mills, K. D., Ferguson, D. O. & Alt, F. W. *Immunol. Rev.* **194**, 77–95 (2003).
3. Raghavan, S. C., Swanson, P. C., Wu, X., Hsieh, C.-L. & Lieber, M. R. *Nature* **428**, 88–93 (2004).
4. Fugmann, S. D., Lee, A. I., Shockett, P. E., Villey, I. J. & Schatz, D. G. *Annu. Rev. Immunol.* **18**, 495–527 (2000).
5. McGuire, E. A. *et al. Mol. Cell. Biol.* **9**, 2124–2132 (1989).
6. Tycko, B. & Sklar, J. *Cancer Cells* **2**, 1–8 (1990).
7. Croce, C. M. & Nowell, P. C. *Adv. Immunol.* **38**, 245–274 (1986).
8. Elliott, B. & Jasin, M. *Cell. Mol. Life Sci.* **59**, 373–385 (2002).
9. Marculescu, R., Le, T., Simon, P., Jaeger, U. & Nadel, B. *J. Exp. Med.* **195**, 85–98 (2002).
10. Chaudhuri, J. *et al. Nature* **422**, 726–730 (2003).

Palaeoclimate

Cool stratification

Roger Francois

The quirky relationship between seawater temperature and density is invoked to account for how, during past global cooling, the high-latitude oceans locked up atmospheric CO₂ and produced further cooling.

Since the beginning of the Tertiary, 65 million years ago, the Earth has experienced a long-term cooling trend¹. The reasons invoked to explain the trend are complex, involving changes in major ocean currents and continental topography, the amount of solar energy reflected by the Earth, concentrations of atmospheric CO₂, and continental glaciations. In contrast, Sigman and colleagues (page 59 of this issue²) highlight a simple yet largely overlooked factor, which may have contributed significantly to cooling during the past 3 million years or so. It is based on the temperature-dependent density of sea water, and the interplay with the other main contributor to density, salinity.

One of the principal determinants of the Earth's mean temperature is the level of atmospheric CO₂, which regulates its radiative balance. One possible contributor to global cooling during the Tertiary is thus a reduction of atmospheric CO₂ concentration with time³. Such a reduction would have resulted from a change in the balance between CO₂ emitted by volcanoes and ocean ridges, and CO₂ uptake by chemical weathering of the continental crust, which between them control the amount of carbon present in the ocean, on continents and in the atmosphere on a timescale of millions of years. A variety of oceanic processes determine the partitioning of CO₂ between atmosphere and ocean, and produce higher-frequency variability in atmospheric CO₂ on timescales of 10⁴ to 10⁵ years. Particularly well documented are cyclical variations, which occurred during the past 500,000 years in concert with variations in Earth's orbital parameters and which resulted in lower atmospheric CO₂ during glacial periods⁴.

One of the leitmotifs that are used to try to explain this lower glacial CO₂ is stratification in the upper ocean at high latitudes. Some 20 years ago, this mechanism was identified as a particularly effective means of sequestering carbon in the deep sea at the expense of the atmosphere⁵. In the modern

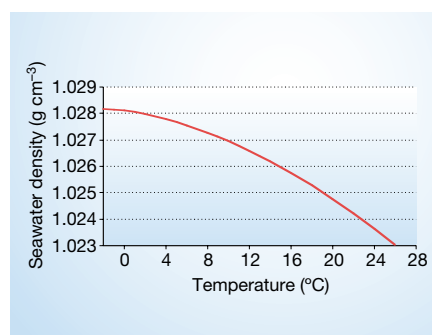


Figure 1 Variation in seawater density as a function of temperature, with salinity at 35‰. Sigman *et al.*² argue that this relationship, in which the density of sea water increases less rapidly as the temperature drops towards freezing point, has profound consequences in partly determining ocean stratification at high latitudes and sequestration of atmospheric CO₂ to the deep sea.

ocean, there is extensive vertical mixing at high latitudes, particularly in the cold waters surrounding Antarctica. This mixing returns to the atmosphere some of the CO₂ that accumulates in the deep sea from the decay of sinking organic matter produced in the sunlit surface waters. Density stratification of the upper ocean in these regions would prevent such return and keep CO₂ sequestered in the deep sea. This finding prompted an extensive search in the sedimentary record for evidence that such stratification happened during the recurring ice ages of the past 500,000 years. Some compelling but not universally accepted evidence for glacial stratification in the Antarctic emerged from this work, but with no clear consensus about the mechanism responsible.

Sigman *et al.*² have also found evidence for enhanced ocean stratification at high latitudes starting 2.7 million years ago, a period coinciding with the initiation of glaciation in the Northern Hemisphere. That evidence comes from the past record of the activity of diatoms, one of the main members of the phytoplankton, whose cell walls are made from opal (a mineral composed of

amorphous silica). Sigman *et al.* report a dramatic decrease in opal accumulation rates in the sediment deposited in the North Pacific and the Antarctic Ocean at that time, which they interpret as a reduction in opal production by diatoms and in the phytoplankton productivity of both regions. They also measured the nitrogen isotopic composition of bulk sediment, and the results suggest that these fewer diatoms used either a higher (in the North Pacific) or a similar (in the Antarctic) fraction of the nitrate nutrients supplied to the sunlit layer by vertical mixing.

Combining these two observations, the authors conclude that nutrient supply to surface waters by vertical mixing must have been reduced, implying increased vertical stability of the upper water column in these two oceanic regions, starting 2.7 million years ago. These trends and their interpretation are similar to those for the Antarctic Ocean during the last glacial period, between 80,000 and 20,000 years ago. Sigman *et al.*, however, go further by proposing a mechanism to account for the stratification.

That mechanism is based on the knowledge that the increase in the density of sea water with decreasing temperature becomes less rapid as the temperature approaches the freezing point (Fig. 1). A stable water column is vertically stratified, with seawater density increasing with depth. At low latitude, where surface waters are warm, there is a strong temperature-controlled density gradient, or thermocline, between depths of about 200 m and 1,000 m, where the temperature drops from above 20 °C to below 5 °C. At high latitudes, where surface-water temperatures are much colder and approach those of deep waters, this thermocline disappears.

Because of that, stratification in cold regions at high latitudes is mainly controlled by another factor, the salinity of surface water, which depends on the balance between precipitation and evaporation, or sea-ice melting and formation. In the modern ocean, only the North Pacific has surface salinities low enough to be permanently stratified. The Antarctic Ocean also has low surface salinity, but not low enough to produce permanent stratification. Instead, there is seasonal stratification during summer melting of the sea ice produced in winter. The North Atlantic Ocean has higher surface salinity, which prevents stratification and promotes deep-water formation. The contrast in surface salinity between the North Pacific and North Atlantic is due to the transport of water vapour from the Atlantic to the Pacific as a result of the interaction between atmospheric circulation and continental topography.

Sigman *et al.*² argue that, before 2.7 million years ago, high-latitude surface waters and the resulting deep waters were sufficiently warm that winter cooling would have overwhelmed any vertical salinity gradient.



100 YEARS AGO

In his notice of my "Papers on Education," in taking exception to my nomenclature, Prof. Smithells has touched on a question of much importance to teachers. "Chalk gas seems unnecessary," he says, "even as a temporary name for carbon dioxide. Why not 'Fixed air,' which is both descriptive and historical?"... "One remark struck me. He does not seem to appreciate that by calling the gas 'Fixed air' you must presuppose that it is fixed and hence all that the word 'Fixed' entails of a knowledge of the gas; whereas, your name is eminently descriptive and entails no knowledge of the gas at all but simply described the source from which it was first obtained."

Henry E. Armstrong

But I think that history usually supplies a good professional name, such as inflammable air, calx of lead, spirit of nitre... To call carbon dioxide chalk-stuff gas asserts that it comes from chalk, or that, in other words, it is a kind of air fixed somehow in chalk... Historically it was called fixed air, and I value that name because Black's clear perception and proof that a gas could be fixed in a solid and be a weighable part of it was the means of inspiring Lavoisier with the right view of the part played by air in the calcination of metals, and so led to results of revolutionary importance.

Arthur Smithells

From *Nature* 3 March 1904.

50 YEARS AGO

In connexion with some recent legal proceedings, a new method for detecting fingerprints has been discovered... The method involves the well-known ninhydrin test for amino-acids, often used in chromatography. In this method, fingerprints on paper have always been considered a great nuisance, and one is often recommended to use forceps "to avoid fingerprints". In our opinion, the method will be most suitable for detecting fingerprints on paper and similar materials... A spontaneous fingerprint contains 98.5–99.5 per cent water, the rest being organic and inorganic compounds. Normally, the water will evaporate, leaving the rest as a fingerprint pattern containing fats, salts, amino-acids, etc. The last group of substances gives the well-known reaction with ninhydrin... Fig. 1 shows part of a page of a French grammar which had not been used for twelve years. The owner's fingerprint can be compared with a fresh print developed on sized paper.

From *Nature* 6 March 1954.

Under these conditions, no permanent stratification would have been possible, and relatively warm, deep water would have formed in most high-latitude regions. As Earth cooled through the Tertiary, however, it reached a threshold coinciding with the onset of glaciation in the Northern Hemisphere. In this situation, high-latitude surface waters became so cold that further cooling in winter did not have enough of an effect on seawater density to overcome the salinity gradients. Surface water would then have been able to accumulate more fresh water by precipitation, exacerbating stratification, which would have then become permanent, as in the modern North Pacific.

Global cooling would then be an important factor promoting high-latitude stratification. This, in turn, would have trapped more carbon from the atmosphere in the deep sea, thereby lowering greenhouse warming and further intensifying cooling. During the Quaternary climatic cycles that followed, interglacial periods would have been sufficiently warm to allow deep winter convection in the Antarctic Ocean, but not in the North Pacific. Presumably, during glacial periods, stratification would also have taken place in the Antarctic⁶, thereby contributing to further lowering of atmospheric CO₂.

A corollary pertinent to present-day concerns is that anthropogenic warming of the

ocean would increase vertical mixing at high latitudes, thereby further increasing atmospheric CO₂ and global change. Although it would take about 1,000 years to fully take effect, this would be a totally uncool legacy for our descendants.

Sigman *et al.*² present compelling evidence for increased high-latitude stratification starting 2.7 million years ago. No doubt their conclusions will stimulate further scrutiny of the history of deep-water temperature⁷, and of the validity of their interpretation of the sedimentary record. Nonetheless, they propose an elegantly simple mechanism that could very well explain some of the major climatic shifts of the geological past. It is refreshing to see that, even in a field as complex as palaeoclimatology, there are still such kernels of simplicity that may be ready for the picking.

Roger Francois is in the Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.

e-mail: rfrancois@whoi.edu

1. Zachos, J. *et al.* *Science* **292**, 686–693 (2001).
2. Sigman, D. M., Jaccard, S. L. & Haug, G. H. *Nature* **428**, 59–63 (2004).
3. Pearson, P. N. & Palmer, M. R. *Nature* **406**, 695–699 (2000).
4. Petit, J. R. *et al.* *Nature* **429**, 429–436 (1999).
5. Sarmiento, J. L. & Toggweiler, J. R. *Nature* **308**, 621–624 (1984).
6. Haug, G. H. *et al.* *Nature* **401**, 779–782 (1999).
7. Lear, C. H. *et al.* *Earth Planet. Sci. Lett.* **210**, 425–436 (2003).

Cell division

Feeling tense enough?

Iain M. Cheeseman and Arshad Desai

Accurately distributing half of each replicated chromosome to both daughters is a major challenge for dividing cells. The mechanisms used to achieve this are becoming apparent, thanks to studies old and new.

At the beginning of mitosis, the process of cell division, chromosomes are organized randomly — like jigsaw puzzle pieces spread out on the floor. Their constituent two 'sister chromatids', each of which contains one of the two identical DNA molecules produced by replication, must be oriented such that they will be pulled in opposite directions into the two newly forming cells. Like a jigsaw, the solution for correctly orienting all chromosomes comes partly through trial and error. Mechanisms must exist to eliminate wrong configurations while selecting the right ones. The nature of such mechanisms was first suggested by micromanipulation experiments begun more than 30 years ago¹. Recent molecular analyses in budding yeast, including a new paper from Dewar *et al.* on page 93 of this issue², extend this earlier work by showing that mechanical tension is crucial to solving the chromosome orientation puzzle.

To segregate chromosomes, eukaryotic cells assemble spindles, bipolar arrays of microtubule polymers formed from the protein tubulin. Spindle microtubules attach to a single site on each chromatid where a large multi-protein complex, the kinetochore, is formed³. When mitosis begins, microtubules start a 'search and capture' process to find these kinetochores. This process is random, so some kinetochores fail to attach efficiently to spindle microtubules. Others generate incorrect attachments — for example, where both sister kinetochores attach to the same pole (Fig. 1). Such 'syntelic' attachments must be corrected so that all chromosomes are bi-oriented — that is, sister kinetochores are stably attached to microtubules from opposite spindle poles. Only then can the glue holding the sister chromatids together be dissolved so that the two spindle poles and their associated chromatids are distributed to the daughter cells.

Micromanipulation studies¹ first provided

evidence that tension between two kinetochores — which is generated only in the bi-oriented state (Fig. 1a) — discriminates between bi-oriented and syntelic attachments. However, these studies were conducted on cells undergoing the first phase of meiosis, the reductive division used to generate gametes. During this phase, two homologous chromosomes, rather than sister chromatids, are connected to opposite spindle poles, so the relevance of this work in mitosis was unclear. Studies of mitotic chromosomes have shown that sister kinetochores face in opposite directions regardless of whether they are attached to spindle microtubules⁴. So when one kinetochore attaches to the spindle it might geometrically force its sister to face the opposite spindle pole and thereby prevent syntely. But recent work demonstrating frequent syntelic attachments during mitosis suggests that additional mechanisms also function to ensure bi-orientation. Consistent with this expectation is the observation that syntelic attachments are corrected by a highly conserved enzyme, Aurora B kinase, in both yeast and vertebrates^{5,6}.

Now, Dewar *et al.*² have elegantly manipulated budding yeast chromosomes to show that the tension generated in the physical linkage between bi-oriented kinetochores and the activity of Aurora B, rather than any specific chromosomal architecture, are sufficient to ensure that the sister chromatids are appropriately aligned. They reached this striking conclusion in two ways, both of which relied on microscopic imaging of fluorescently tagged chromosomes to study their bi-orientation in living cells.

Under normal circumstances, protein complexes termed cohesins hold sister chromatids together until all chromosomes are properly aligned on the spindle. In the bi-oriented state, microtubules from opposite spindle poles attach to sister kinetochores, and poleward forces act on the kinetochores to generate tension in the connection between them. Using yeast cells that contained defective cohesin and inactive topoisomerase II, an enzyme that resolves physically intertwined DNA, Dewar *et al.*² showed that sister chromatids with physically linked DNA molecules can efficiently bi-orient without cohesin. A similar finding was also reported recently in vertebrate cells⁷. So cohesin complexes and the specific connection they form between sister chromatids are not required for bi-orientation — a different type of physical connection will suffice.

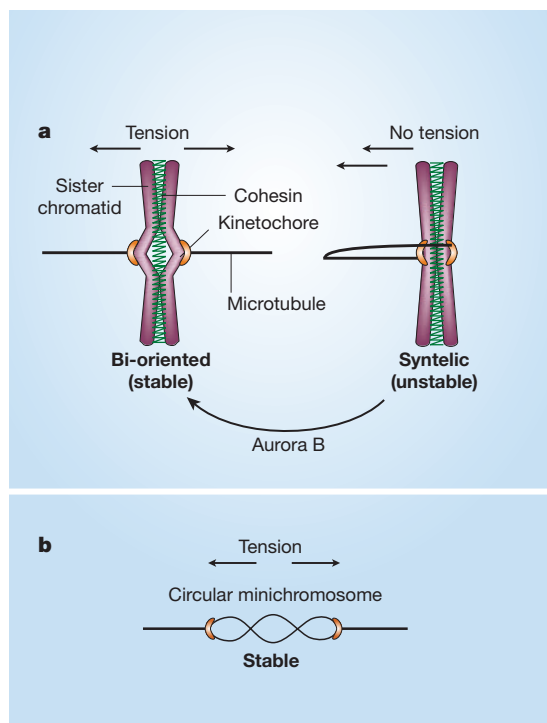


Figure 1 Bi-orientation is required for accurate division of sister chromatids. **a**, Microtubules of the spindle attach to sister chromatids through multi-protein complexes known as kinetochores. If the microtubules are connected to the opposing spindle poles, the chromosomes are said to be bi-oriented, and tension is generated because the sister chromatids are held together by a cohesin glue. If, however, the kinetochores of sister chromatids attach to microtubules coming from the same pole, the chromosomes are said to be syntelic, and no tension is generated. The Aurora B kinase recognizes and corrects the lack of tension in syntelic chromosomes. **b**, Dewar *et al.*² made a circular minichromosome containing two kinetochores and showed that tension could still be generated in the absence of cohesin because the kinetochores were still physically connected.

In a technical *tour de force*, the authors engineered a circular minichromosome with two kinetochores on a single DNA molecule. Remarkably, these artificial minichromosomes bi-orient with similar efficiency and kinetics to those of normal chromosomes (Fig. 1b). Thus, any physical connection between two kinetochores that supports the development of tension can facilitate bi-orientation.

Dewar *et al.*² also show that Aurora B kinase is required to bi-orient the artificial minichromosomes with two kinetochores. Therefore, Aurora B must recognize incorrect attachments simply by virtue of their lack of tension, regardless of the precise structure that links the two kinetochores. Finding out how Aurora B identifies and corrects them is an obvious next step. In budding yeast, where kinetochores assemble on a short (120-base-pair) piece of DNA and bind to just a single microtubule, Aurora B must periodically detach kinetochores from microtubules until the tense, bi-oriented

state is achieved. The authors² provide direct evidence for this by generating yeast cells that have more than two spindle poles. In such aberrant cells, an engineered minichromosome with just one kinetochore, which can only form a single attachment and thus lacks tension, moves repeatedly between the different spindle poles as though it is searching for a stable conformation. If Aurora B function is inhibited, the minichromosome remains linked to just one spindle pole, indicating that Aurora B detaches microtubules from kinetochores in the absence of tension.

In contrast to budding yeast, kinetochores of other eukaryotes bind multiple microtubules (about 20 in humans)⁸. These larger kinetochores must coordinate all these microtubules and also deal with incorrect attachments in which microtubules from opposite spindle poles connect to a single kinetochore (termed 'merotelly')⁹. Another study¹⁰, in this month's *Nature Cell Biology*, found that Aurora B does not merely detach syntelic kinetochores from microtubules in vertebrates — it orchestrates the coordinated disassembly of all the microtubules that are bound to each kinetochore, so that the syntelically oriented chromosomes move towards the spindle poles before they are bi-oriented.

Although sister kinetochore geometry seems to be dispensable in budding yeasts with their single-microtubule-connected kinetochores, it could contribute to reducing merotelly, as implied by the conservation of this aspect of chromosome architecture throughout eukaryotic evolution. Tackling the extra dimension that the multiplicity of microtubule-binding sites at kinetochores introduces will undoubtedly be another brain-teaser — and a particularly important one, too, because the loss of a single chromosome can be lethal, and aberrant numbers of chromosomes can contribute to birth defects and cancer.

Iain M. Cheeseman and Arshad Desai are at the Ludwig Institute for Cancer Research and the University of California, San Diego, La Jolla, California 92093, USA.

e-mails: icheeseman@ucsd.edu
abdesai@ucsd.edu

1. Nicklas, R. B. & Koch, C. A. *J. Cell Biol.* **43**, 40–50 (1969).
2. Dewar, H., Tanaka, K., Nasmyth, K. & Tanaka, T. U. *Nature* **428**, 93–97 (2004).
3. Cleveland, D. W., Mao, Y. & Sullivan, K. F. *Cell* **112**, 407–421 (2003).
4. Rieder, C. L. *Int. Rev. Cytol.* **79**, 1–58 (1982).
5. Tanaka, T. U. *et al.* *Cell* **108**, 317–329 (2002).
6. Hauf, S. *et al.* *J. Cell Biol.* **161**, 281–294 (2003).
7. Vagnarelli, P. *et al.* *EMBO Rep.* **5**, 167–171 (2004).
8. Bloom, K. *Cell* **73**, 621–624 (1993).
9. Cimmini, D. *et al.* *J. Cell Biol.* **153**, 517–527 (2001).
10. Lampson, M. A., Renduchitala, K., Khodjakov, A. & Kapoor, T. M. *Nature Cell Biol.* **6**, 232–237 (2004).

Complex systems

Predicting the unpredictable

Phys. Rev. Lett. **92**, 074105 (2004)

Some processes seem impossible to predict without actually letting them unfold and seeing what happens — they are ‘computationally irreducible’. That appears to undermine the entire mission of science, which aims to understand and predict rather than merely to mimic. But all is not lost, say Navot Israeli and Nigel Goldenfeld. They show that at least some such processes can be rendered predictable by coarse-graining — in essence, by not worrying too much about the fine details.

They have looked at computationally irreducible cellular automata (CA). These are systems composed of many components distributed on a lattice, which switch between several possible states according to local rules governing the interactions between cells. CA are used to model many complex dynamical processes, from ecosystem evolution to material fracture and biological growth. Stephen Wolfram has proposed that many of the CA that evolve in complex patterns are computationally irreducible, and so are inherently unpredictable. But Israeli and Goldenfeld show that by sacrificing some level of detail (as real experiments must always do), most such CA can be reduced to computationally reducible ones. A theoretical model of the systems’ behaviour then becomes feasible, introducing some predictability of their future evolution.

Philip Ball

Apoptosis

Danse macabre

Neuron **41**, 535–547 (2004)

During normal brain development many neurons die — one way or another, these cells fall short of the high standards required. The neurons die autonomously, in an orderly manner, and their remains are removed instantly. Or so the theory went, until José Luis Marín-Teva and colleagues recently provided evidence that microglia cells, once thought to be innocent corpse collectors, may actually deliver the final blow to cells on death row.

Microglia engulf cells, moving about in the central nervous system and removing waste, including other cells. Specific signals on dying cells can flag them down. Usually, an engulfing cell binds the dying cell, and — as if embracing it — extends its membranes around the corpse and eats it.

But studies in the nematode worm had already indicated that not all engulfing cells are as innocent as they seem. When Marín-Teva *et al.* selectively removed microglia from slice preparations of

postnatal mouse cerebella, numerous Purkinje neurons that would otherwise have died within 24 hours *in vitro* were reprimed. This suggested that microglia actually help the Purkinje cells to die. The authors also propose that the release of superoxide ions, which are generated during typical respiratory bursts in the engulfing cells, may cause the Purkinje cells to die ‘in the arms’ of microglia.

Marie-Thérèse Heemels

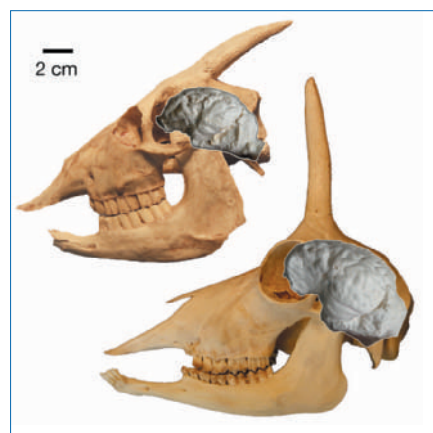
Palaeobiology

Small brainer

Brain Behav. Evol. **63**, 125–140 (2004)

Mammals have tended to acquire bigger brains during their evolution. But Meike Köhler and Salvador Moyà-Solà find that this was apparently not the case for a goat-like animal, *Myotragus*, which lived in Majorca from about 5 million years ago.

From measurements of fossils, they show that the *Myotragus* brain was much smaller than that of living ruminants of the same weight (see picture), whereas the brain size of a close relative that lived on the French mainland matched that of modern-day species.



Isolated effects? Bottom, the skull and brain size of a living ruminant of the same body size as (top) the Majorcan *Myotragus*.

Köhler and Moyà-Solà invoke geographical isolation as an explanation. According to the fossil evidence, the ancestor of *Myotragus* settled in Majorca at a time when the Mediterranean Sea had largely dried out and islands were joined to the mainland. These ancestral goats would have been exposed to an array of predators and had a variety of food resources.

When the sea level rose, however, Majorca was left without large predators. The authors argue that, given the need to conserve energy, of which the brain is a large consumer, and the reduced requirement for fast-acting senses and motor responses to see off predators, a smaller brain was a predictable evolutionary upshot.

Lucy Odling-Smee

Astronomy

Beryllium clocks on

Astrophys. J. **602**, 528–542 (2004)

Meteorites that formed roughly 4.5 billion years ago carry clues about the birth of our Solar System. They once contained radioactive beryllium-10 (^{10}Be), which is created when high-energy particles knock protons and neutrons out of carbon, nitrogen or oxygen atoms. Unlike other elements found in the same meteorites, ^{10}Be is not formed in the fusion reactions of stars or supernovae.

The ^{10}Be content of meteorites has led to speculation that radiation from the proto-Sun was the driving force in the element’s synthesis. If this were the case, other elements (such as ^{26}Al) could have a similar origin, rather than being injected into the early solar nebula by a supernova explosion (as proposed in a leading theory on the genesis of the Solar System).

By modelling the flux of energetic particles in the early solar nebula, S. J. Desch *et al.* have calculated that galactic cosmic rays alone could be enough to explain the high levels of ^{10}Be , without having to invoke solar radiation. This would mean that the ratio of ^{10}Be to the stable isotope, ^9Be , would initially have been the same for all solid material in the solar nebula. Any variation in this ratio could therefore provide a ‘beryllium chronometer’ to date extreme events in the early Solar System.

Mark Peplow

Neuroscience

Stem cells enjoy long life

Exp. Cell Res. doi:10.1016/j.yexcr.2003.11.025 (2004)

Like many other cell types, human neural stem cells only divide a finite number of times in culture. This cellular ageing has generally been put down to telomere erosion — the tips of chromosomes become progressively shorter with every round of DNA replication until the cells are no longer viable.

Now Ana Villa *et al.* describe the characteristics of a human neural stem-cell line that remains youthful. The cells, which have been growing for four years, were immortalized by the addition of the growth-promoting gene *v-myc*. The gene appears to counter the effects of progressive telomere shortening, so that although the telomeres of the cells were short, they were stable. *v-myc* also boosted levels of telomerase, an enzyme that helps keep telomeres long. Importantly, another stem-cell-like quality — the ability to generate mature cell types — was retained, and as the stem cells developed into neurons, *v-myc* expression and telomerase activity decreased.

The study provides a controlled method for the long-term culture of human neural stem cells.

Helen R. Pilcher

Ant parasite queens revert to mating singly

A parasitic ant has abandoned the multiple mating habit of the queens of its related host.

Multiple mating (polyandry) is widespread among animal groups, particularly insects¹. But the factors that maintain it and underlie its evolution are hard to verify because benefits and costs are not easily quantified and they tend to be similar in related species. Here we compare the mating strategies of the leaf-cutting ant *Acromyrmex echinator* and its recently derived social parasite *Acromyrmex insinuator*, which is also its closest relative² (see Fig. 1). We find that although the host queens mate with up to a dozen different males, the social parasite mates only singly. This rapid and surprising reversion to single mating in a socially parasitic ant indicates that the costs of polyandry are probably specific to a free-living lifestyle.

In social insects, obligate polyandry has evolved independently in four groups (honeybees, vespid wasps, harvester ants and leaf-cutting ants³), among which every species was previously believed to be multiply mated. All free-living leaf-cutting ants are multiply mated⁴, so the common ancestor of both parasite and host ants must have had multiply mated queens. Because social parasite queens exploit the resources and workers of a host colony to produce reproductive offspring (sexuals) without investing in a large worker force of their own^{5–7}, we investigated whether this dramatic change in lifestyle could have altered the costs and benefits of polyandry, and thereby driven a shift in mating strategy.

Parasitized *A. echinator* colonies were collected from Gamboa, Panama, where the species is common⁷. We compared the mating frequencies of host and parasite by analysing the genotypes of queens and their female offspring to estimate the number of fathers contributing to a queen's progeny. Host queens mated with 10.6 ± 1.0 males (mean $\pm$ s.e.; $n=10$) but social parasite queens were predominantly singly mated (average number of matings per queen, 1.19 ± 0.10 ; only 2 out of 13 queens doubly mated; see Fig. 2).

This difference is statistically highly significant (Mann–Whitney U -test, $P < 0.0001$) and robust to alternative measures of queen-mating frequency. For example, the observed number of queen matings was 9.3 ± 0.7 versus 1.15 ± 0.10 ($P < 0.0001$), the effective number of queen matings was 5.3 ± 0.6 versus 1.09 ± 0.09 ($P < 0.0001$), the estimated proportion of multiply mated queens was 1.00 (95% confidence interval, 0.69–1.00) versus 0.19 (0.034–0.51; binomial test,



Figure 1 A multiply mated queen of the leaf-cutting ant *Acromyrmex echinator* (right) and its singly mated socially parasitic companion queen *A. insinuator* (left) on their joint fungus garden. Small worker ants cluster around the parasite queen.

$P = 0.0006$), and genetic relatedness among the offspring was 0.380 ± 0.042 versus 0.693 ± 0.053 (t -test, $P = 0.0002$) for the host and parasite, respectively. The difference cannot be attributed to low variability in the genetic markers⁸, sampling error⁸ or inbreeding: neither species was inbred ($F = 0.006$, $n = 10$, $P = 0.19$ and $F = -0.035$, $n = 19$, $P = 0.88$ for *A. echinator* and *A. insinuator*, respectively).

These findings provide insight into the selective forces affecting the evolution of polyandry. Like its host, *A. insinuator* is outbred and produces hundreds of winged sexuals in discrete batches that leave their natal colonies within the same brief time period⁷. This indicates that females of both species mate with unrelated males in swarms outside the colony. The parasite is also very abundant, infecting 44% of host colonies at our study site⁷. Mating behaviour, mate availability and thus the costs of mating are therefore likely to be similar for the two species. That *A. insinuator* has reverted to single mating suggests that the benefits of polyandry are reduced for the social parasite. The reversion to single mating seems to have been rapid compared with the evolution of other adaptive traits^{5–7}, supporting the widely held but rarely demonstrated assumption^{1,2} that multiple mating is costly.

The association between a reversion to single mating and a switch to social parasitism indicates that the benefits of polyandry for *Acromyrmex* ants may be specific to being free-living and social. Four proposed benefits of multiple mating for social insect queens (nutrient transfer, sperm competition, variable daughter queens and avoidance of genetic incompatibility³) are unlikely to be important in the maintenance of polyandry in *Acromyrmex*, because they

would apply equally to social parasites and free-living species. Other likely benefits (greater sperm store, convergence of sex-allocation interests between queens and workers, reduced fitness cost of producing sterile diploid males and increased genetic diversity among workers³) are specific to free-living social insects and so are the most likely explanations for polyandry in *Acromyrmex* ants. The benefits of polyandry in these ants are therefore related to their social lifestyle, indicating that the selective forces affecting the evolution of polyandry in free-living social insects differ from those in other animals. We would expect workerless social parasites of other polyandrous social insects to have also reverted to

mating singly. The identification of additional reversions to single mating, both in social and non-social animals, will be important in further distinguishing the factors

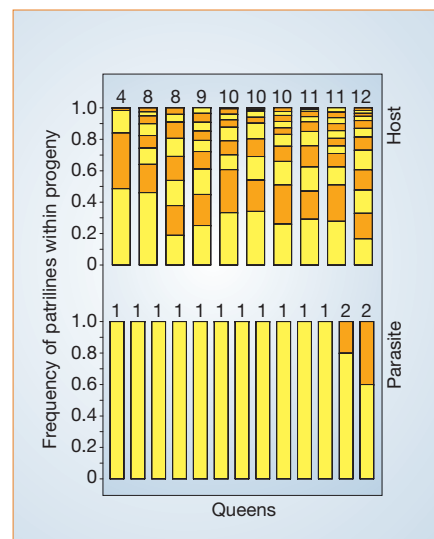


Figure 2 The observed frequency distributions of patrilineal inheritance (offspring sired by different males) of the social parasite *Acromyrmex insinuator* ($n=13$) and its host *A. echinator* ($n=10$), as derived from genetic analysis of the progeny of single queens. Patrilineal inheritance is shown by alternate shading patterns, with their total number in each colony given above the bars. Host queens and an average of 203 (range, 35–672) female offspring per queen were genotyped at four microsatellite loci (*Ech1390*, *Ech3385*, *Ech4126* and *Ech4225*)⁹. Parasite queens and an average of 14.7 (6–20) female progeny per queen were genotyped at two loci (*Trachy11_12* and *Ech3385*)^{9,10}. Genotypes of queens and their offspring were analysed using the software MATESOFT 1.0b (by A. Moilanen, L. Sundström and J. S. P.; available at <http://www.zi.ku.dk/popecol/personal/JSPedersen/MateSoft.htm>) to obtain mating-frequency statistics⁸, and the total mate number for host queens was further estimated using a modified version of the coverage statistics calculated by the program CHAO¹¹.

involved in the evolution and maintenance of multiple mating by females.

Seirian Sumner*, William O. H. Hughes, Jes S. Pedersen, Jacobus J. Boomsma

Biological Institute, Department of Population Biology, University of Copenhagen, 2100 Copenhagen, Denmark

e-mail: jboomsma@zi.ku.dk

*Present addresses: Smithsonian Tropical Research Institute, Unit 0948, APO AA 34002, USA, and Institute of Zoology, Zoological Society of London, Regents Park, London N1 4RY, UK.

1. Simmons, L. W. *Sperm Competition and its Evolutionary Consequences in the Insects* (Princeton University Press, Princeton, 2001).
2. Sumner, S., Aanen, D. K., Delabie, J. & Boomsma, J. J. *Insectes Soc.* **51**, 37–42 (2004).
3. Crozier, R. H. & Fjerdingstad, E. J. *Ann. Zool. Fenn.* **38**, 267–285 (2001).
4. Villesen, P., Murakami, T., Schultz, T. R. & Boomsma, J. J. *Proc. R. Soc. Lond. B* **269**, 1541–1548 (2002).
5. Sumner, S., Nash, D. R. & Boomsma, J. J. *Proc. R. Soc. Lond. B* **270**, 1315–1322 (2003).
6. Sumner, S., Hughes, W. O. H. & Boomsma, J. J. *Behav. Ecol. Sociobiol.* **54**, 256–263 (2003).
7. Bekkevold, D. & Boomsma, J. J. *J. Evol. Biol.* **13**, 615–623 (2000).
8. Pedersen, J. S. & Boomsma, J. J. *Mol. Ecol.* **8**, 577–587 (1999).
9. Ortius-Lechner, D., Gertsch, P. J. & Boomsma, J. J. *Mol. Ecol.* **9**, 114–116 (2000).
10. Villesen, P., Gertsch, P. J. & Boomsma, J. J. *Mol. Ecol. Notes* **2**, 320–322 (2002).
11. Chao, A. & Lee, S. M. J. *Am. Stat. Assoc.* **87**, 210–217 (1992).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Arctic Ocean

Hydrothermal activity on Gakkel Ridge

In the hydrothermal circulation at mid-ocean ridges, sea water penetrates the fractured crust, becomes heated by its proximity to the hot magma, and returns to the sea floor as hot fluids enriched in various chemical elements. In contradiction to earlier results¹ that predict diminishing hydrothermal activity with decreasing spreading rate, a survey of the ultra-slowly spreading Gakkel Ridge (Arctic Ocean) by Edmonds *et al.*² and Michael *et al.*³ suggests that, instead of being rare, the hydrothermal activity is abundant — exceeding by at least a factor of two to three³ what would be expected by extrapolation from observation on faster spreading ridges. Here we use helium-3 (³He), a hydrothermal tracer⁴, to show that this abundance of venting sites does not translate, as would be expected, into an anomalous hydrothermal ³He output from the ridge. Because of the wide implications of the submarine hydrothermal processes for mantle heat and mass fluxes to the ocean, these conflicting results call for clarification of the link between hydrothermal activity and crustal production at mid-ocean ridges.

On the basis of the axial distribution of light scattering and manganese tracer anomalies measured during the Arctic mid-ocean ridge experiment (AMORE), hydrothermal activity on Gakkel Ridge has been reported to be “far more abundant than predicted on

the basis of previous correlations with spreading rate”³. Owing to the high enrichment of the hydrothermal fluids in mantle ³He (typically a factor of 10⁴ relative to ambient sea water), this abundant hydrothermal activity should logically correspond to an enhanced ³He output from the ridge.

At a global scale, three-dimensional oceanic general circulation models make a good job of simulating the oceanic distribution of hydrothermal ³He by using the simple assumption that the ³He source intensity is linearly correlated with the spreading rate^{5,6}. The slope of the ³He flux–spreading rate relationship is close to 330 ³He mol km^{−2} of crustal spreading, translating into a world average ³He annual flux of 0.33 mmol km^{−2} of ridge and per mm yr^{−1} of full spreading rate. Available ³He fluxes determined for active portions of ridge show results consistent with this global trend (Fig. 1a).

Gakkel Ridge stretches 1,800 km across the Eurasian basin of the Arctic Ocean. Residence time of deep and bottom waters in the Greenland/Norwegian seas and the Eurasian basin have been studied in detail in the recent past using multi-tracer balances⁷. ³He concentrations are relatively low in the deep waters of the Arctic Ocean: Eurasian basin deep waters (EBDW), which are at 1,500–2,600 m depth, show a ³He excess relative to solubility equilibrium of 4.3 ± 0.8‰, whereas the Eurasian basin bottom waters (EBBW) average ³He excess is 3.7 ± 0.6‰ (refs 7, 8).

The ³He input from Gakkel Ridge to the EBBW and EBDW can be calculated using a simplified version of a box model⁷ (Fig. 1b). Water fluxes are derived from the ³⁹Ar radioactive tracer budget (³⁹Ar is a cosmogenic isotope with a half-life of 269 years), allowing the mantle ³He flux to be calculated from the ³He budget. Taking into account the uncertainties in the data, a maximum ³He input of 3.4 mol yr^{−1} is inferred. Scaled to the length and spreading rate of Gakkel Ridge, the mantle ³He input translates into an annual flux of 0.15–0.32 mmol km^{−1} of ridge and per mm yr^{−1} of full spreading rate (Fig. 1a), showing no positive anomaly with respect to that expected from the value of the spreading rate.

One explanation mentioned by Michael *et al.*³ for the observed discrepancy could be that the high plume incidence observed during the survey (80% plume incidence on the 145 profiles) overestimated the abundance of vents because of the confined nature of the deep unsegmented rift valley. It is possible that plume meanders originating from a single vent site may have been interpreted as corresponding to different sources. Alternatively, if the inferred number of vent sites is right, an explanation may rest on the fact that vent frequency is not a direct measurement of hydrothermal activity.

It is also possible that, on ultra-slowly spreading ridges, individual plumes carry

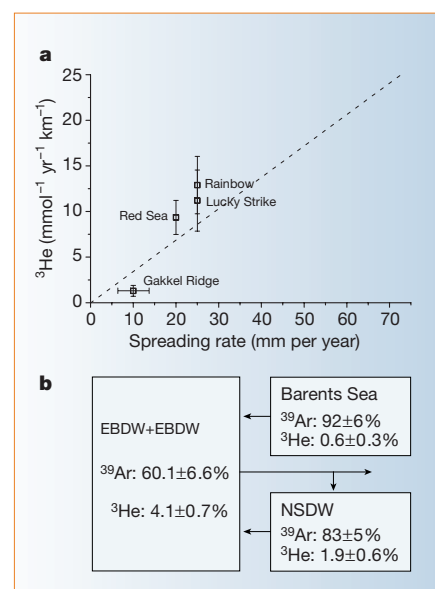


Figure 1 ³He fluxes at mid-ocean ridges. **a**, ³He annual flux per kilometre of ridge as a function of spreading rate. The linear relationship used in GCM simulations^{5,6} is shown (dashed line). The Mid-Atlantic Ridge (MAR) Lucky Strike site and Red Sea values are from refs 9, 10. The Rainbow site (MAR) value is unpublished data. **b**, Eurasian basin box model: Eurasian basin deep and bottom waters (EBDW and EBBW respectively) are directly ventilated by the sinking of dense shelf waters from the Barents Sea. They leave the basin through the western part of the Fram Strait and mix with the Greenland Sea and Norwegian Sea deep waters (NSDW). In addition to the Barents Sea waters, some NSDW also return to the north and re-enter the deep Arctic through the eastern part of the Fram Strait⁷. ³⁹Ar values⁷ are expressed as a percentage of modern concentration. ³He data are expressed using the isotope δ notation.

fewer chemicals and less heat, so higher frequency does not necessarily mean higher hydrothermal activity. The apparent discrepancy between the enhancement factor of two to three in the vent site frequency deduced from along axis tracer surveys and the low ³He flux may therefore primarily reflect the unusual characteristics of the hydrothermal circulation on this ultra-slowly spreading ridge.

Detailed investigation of individual venting sites in such environments is needed to resolve this issue and to reassess the geological parameters that control ridge processes, including hydrothermal plume and vent frequency, fluid circulation and fluxes.

Philippe Jean-Baptiste, Elise Fourné

Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CEA-CNRS, Centre d'Etudes de Saclay, 91191 Gif-sur-Yvette, France
e-mail: pjb@lscs.saclay.cea.fr

1. Baker, E. T., Chen, Y. J. & Phipps Morgan, J. *Earth Planet. Sci. Lett.* **142**, 137–145 (1996).
2. Edmonds, H. N. *et al. Nature* **421**, 252–256 (2003).
3. Michael, P. J. *et al. Nature* **423**, 956–961 (2003).
4. Craig, H. & Lupton, J. E. in *The Sea* Vol. 7, 391–428 (Wiley, New York, 1981).
5. Farley, K. A., Maier-Reimer, E., Schlosser, P. & Broecker, W. S. *J. Geophys. Res.* **100**, 3829–3839 (1995).
6. Dutay, J. C. *et al. J. Mar. Syst.* (in the press).
7. Boenisch, G. & Schlosser, P. *Prog. Oceanogr.* **35**, 29–52 (1995).
8. Schlosser, P. *et al. Prog. Oceanogr.* **35**, 1–28 (1995).
9. Jean-Baptiste, P. *et al. Earth Planet. Sci. Lett.* **157**, 69–77 (1998).
10. Jean-Baptiste, P., Fourné, E., Metz, N., Ternon, J. F. & Poisson, A. *J. Mar. Syst.* (in the press).

Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson¹, Jarrod Chapman^{3,4}, Philip Hugenholtz¹, Eric E. Allen¹, Rachna J. Ram¹, Paul M. Richardson⁴, Victor V. Solovyev⁴, Edward M. Rubin⁴, Daniel S. Rokhsar^{3,4} & Jillian F. Banfield^{1,2}

¹Department of Environmental Science, Policy and Management, ²Department of Earth and Planetary Sciences, and ³Department of Physics, University of California, Berkeley, California 94720, USA

⁴Joint Genome Institute, Walnut Creek, California 94598, USA

Microbial communities are vital in the functioning of all ecosystems; however, most microorganisms are uncultivated, and their roles in natural systems are unclear. Here, using random shotgun sequencing of DNA from a natural acidophilic biofilm, we report reconstruction of near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II, and partial recovery of three other genomes. This was possible because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low. Because each sequence read came from a different individual, we could determine that single-nucleotide polymorphisms are the predominant form of heterogeneity at the strain level. The *Leptospirillum* group II genome had remarkably few nucleotide polymorphisms, despite the existence of low-abundance variants. The *Ferroplasma* type II genome seems to be a composite from three ancestral strains that have undergone homologous recombination to form a large population of mosaic genomes. Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies in an extreme environment.

The study of microbial evolution and ecology has been revolutionized by DNA sequencing and analysis^{1–3}. However, isolates have been the main source of sequence data, and only a small fraction of microorganisms have been cultivated^{4–6}. Consequently, focus has shifted towards the analysis of uncultivated microorganisms via cloning of conserved genes⁵ and genome fragments directly from the environment^{7–9}. To date, only a small fraction of genes have been recovered from individual environments, limiting the analysis of microbial communities as networks characterized by symbioses, competition and partitioning of community-essential roles. Comprehensive genomic data would resolve organism-specific pathways and provide insights into population structure, speciation and evolution. So far, sequencing of whole communities has not been practical because most communities comprise hundreds to thousands of species¹⁰.

Acid mine drainage (AMD) is a worldwide environmental problem that arises largely from microbial activity¹¹. Here, we focused on a low-complexity AMD microbial biofilm growing hundreds of feet underground within a pyrite (FeS₂) ore body^{12–15}. This represents a self-contained biogeochemical system characterized by tight coupling between microbial iron oxidation and acidification due to pyrite dissolution^{11,16,17}. Random shotgun sequencing of DNA from entire microbial communities is one approach for the recovery of the gene complement of uncultivated organisms, and for determining the degree of variability within populations at the genome level. We used random shotgun sequencing of the biofilm to obtain the first reconstruction of multiple genomes directly from a natural sample. The results provide novel insights into community structure, and reveal the strategies that underpin microbial activity in this environment.

Initial characterization of the biofilm

Biofilms growing on the surface of flowing AMD in the five-way region of the Richmond mine at Iron Mountain, California¹², were sampled in March 2000. Screening using group-specific¹⁸

fluorescence *in situ* hybridization (FISH) revealed that all biofilms contained mixtures of bacteria (*Leptospirillum*, *Sulfobacillus* and, in a few cases, *Acidimicrobium*) and archaea (*Ferroplasma* and other members of the Thermoplasmatales). The genome of one of these archaea, *Ferroplasma acidarmanus* fer1, isolated from the Richmond mine, has been sequenced previously (http://www.jgi.doe.gov/JGI_microbial/html/ferroplasma/ferro_homepage.html).

A pink biofilm (Fig. 1a) typical of AMD communities was selected for detailed genomic characterization (see Supplementary Information). The biofilm was dominated by *Leptospirillum* species and contained *F. acidarmanus* at a relatively low abundance (Fig. 1b, c). This biofilm was growing in pH 0.83, 42 °C, 317 mM Fe, 14 mM Zn, 4 mM Cu and 2 mM As solution, and was collected from a surface area of approximately 0.05 m².

A 16S ribosomal RNA gene clone library was constructed from DNA extracted from the pink biofilm, and 384 clones were end-sequenced (see Supplementary Information). Results indicated the presence of three bacterial and three archaeal lineages. The most abundant clones are close relatives of *L. ferriphilum*¹⁹ and belong to *Leptospirillum* group II (ref. 13). Although 94% of the *Leptospirillum* group II clones were identical, 17 minor variants were detected with up to 1.2% 16S rRNA gene-sequence divergence from the dominant type. Tightly defined groups (up to 1% sequence divergence) related to *Leptospirillum* group III (ref. 13), *Sulfobacillus*, *Ferroplasma* (some identical to fer1), 'A-plasma'¹⁵ and 'G-plasma'¹⁵ were also detected. *Leptospirillum* group III, G-plasma and A-plasma have only recently been detected in culture-independent molecular surveys. FISH-based quantification (Fig. 1c; see also Supplementary Information) confirmed the dominance of *Leptospirillum* group II in the biofilm.

Community genome sequencing and assembly

In conventional shotgun sequencing projects of microbial isolates, all shotgun fragments are derived from clones of the same genome. When using the shotgun sequencing approach on genomes from an

environmental sample, however, variation within each species population might complicate assembly. If intraspecies variation is dominated by limited local polymorphism or homologous recombination, it should be possible to define a composite genome for each species population. Conversely, if the genomic heterogeneity within a species is dominated by large rearrangements, deletions, or insertions, it may be impossible to define composite genomes for species populations from natural communities.

A small insert plasmid library (average insert size 3.2 kilobases (kb)) was constructed from the biofilm DNA for random shotgun sequencing (see Supplementary Information). A total of 76.2 million base pairs (bp) of DNA sequence was generated from 103,462 high-quality reads (averaging 737 bp per read). Analysis of raw shotgun data (Supplementary Figs S1–5) indicated the presence of both bacterial and archaeal genomes at sequence coverages of up to 10 $\times$, which would be sufficient to produce a high-quality assembly from a conventional microbial genome project^{20,21}. The shotgun data set was assembled with JAZZ, a whole-genome shotgun assembler²². Anticipating polymorphisms, we permitted alignment discrepancies beyond those expected from sequencing error if they were consistent with end-pairing constraints. Over 85% of the shotgun reads were assembled into scaffolds longer than 2 kb (a scaffold is a reconstructed genomic region that may contain gaps of a known size range). The combined length of the 1,183 scaffolds is 10.83 megabases (Mb). The assembly is internally self consistent, with 97.2% of end pairs from the same clone assembled with the appropriate orientation and separation, as expected for a low rate of mispairing error (tracking and chimaeric clones).

The first step in assignment of scaffolds to organism types was to

separate the scaffolds by average G+C content. These were subsequently subdivided using read depth (coverage). Dinucleotide frequencies did not allow for further subdivision. Notably, separation of scaffolds into low G+C (<43.5%; Supplementary Fig. S3a) and high G+C ($\geq$ 43.5%) content ‘bins’ was not significantly compromised by local heterogeneities in G+C content because the scaffolds were binned after assembly. As the scaffolds are typically tens of kilobases long, local fluctuations in G+C content are averaged over the length of each scaffold, allowing, in most cases (>99%), clear assignment to bins of high or low G+C content.

The high G+C scaffolds at approximately 10 $\times$ coverage (70 scaffolds up to 137 kb in length, totalling 2.23 Mb) were identified by the presence of a single 16S rRNA gene as belonging to the genome of a *Leptospirillum* group II species. The average G+C content (55.8%) is comparable to the G+C content (54.9–58%) of *L. ferrophilum*¹⁹. The total high G+C scaffold length is close to the estimated genome size of *Leptospirillum ferrooxidans*²³ (1.9 Mb). This suggests that essentially the entire *Leptospirillum* group II genome was recovered from the community DNA.

The low G+C scaffolds at approximately 10 $\times$ coverage were assembled into 59 scaffolds of up to 138 kb in length, totalling 1.82 Mb. The single 16S rRNA gene identified in these scaffolds was 99% identical to that of the fer1 isolate; however, alignment of the scaffolds to the fer1 genome revealed an average of 22% divergence at the nucleotide level (Supplementary Fig. S6). The total scaffold length is close to the genome size of fer1 (1.9 Mb; Allen *et al.*, unpublished data), and local gene order and content are highly conserved (Supplementary Fig. S7). Therefore, these 59 scaffolds represent a nearly complete genome of a previously unknown, uncultured *Ferroplasma* species distinct from fer1. We designate this as *Ferroplasma* type II. The dominance of this organism type was unexpected before the genomic analysis.

We assigned the roughly 3 $\times$ coverage, high G+C scaffolds to *Leptospirillum* group III on the basis of rRNA markers (474 scaffolds up to 31 kb, totalling 2.66 Mb). Comparison of these scaffolds with those assigned to *Leptospirillum* group II indicates significant sequence divergence and only locally conserved gene order, confirming that the scaffolds belong to a relatively distant relative of *Leptospirillum* group II. A partial 16S rRNA gene sequence from *Sulfobacillus thermosulfidooxidans* was identified in the unassembled reads, suggesting very low coverage of this organism. If any *Sulfobacillus* scaffolds >2 kb were assembled, they would be grouped with the *Leptospirillum* group III scaffolds.

We compared the 3 $\times$ coverage, low G+C scaffolds (580 scaffolds, 4.12 Mb) to the fer1 genome in order to assign them to organism types (Supplementary Fig. S6). Scaffolds with $\geq$ 96% nucleotide identity to fer1 were assigned to an environmental *Ferroplasma* type I genome (170 scaffolds up to 47 kb in length and comprising 1.48 Mb of sequence). The remaining low-coverage, low G+C scaffolds are tentatively assigned to G-plasma. The largest scaffold in this bin (62 kb) contains the G-plasma 16S rRNA gene. The 410 scaffolds assigned to G-plasma comprise 2.65 Mb of sequence. A partial 16S rRNA gene sequence from A-plasma was identified in the unassembled reads, suggesting low coverage of this organism. Any scaffolds from A-plasma >2 kb would be included in the G-plasma bin. Although eukaryotes are present in the AMD system, they were in low abundance in the biofilm studied. So far, no scaffolds from eukaryotes have been detected.

As independent evidence that the *Leptospirillum* group II and *Ferroplasma* type II genomes are nearly complete, we located a full complement of transfer RNA synthetases in each genome data set. An almost complete set of these genes was also recovered from *Leptospirillum* group III. The G-plasma bin contains more than a full set of tRNA synthetases, consistent with inclusion of some A-plasma scaffolds. In addition, we established that the *Leptospirillum* group II, *Leptospirillum* group III, *Ferroplasma* type I, *Ferroplasma* type II and G-plasma bins contained only one set of rRNA genes.

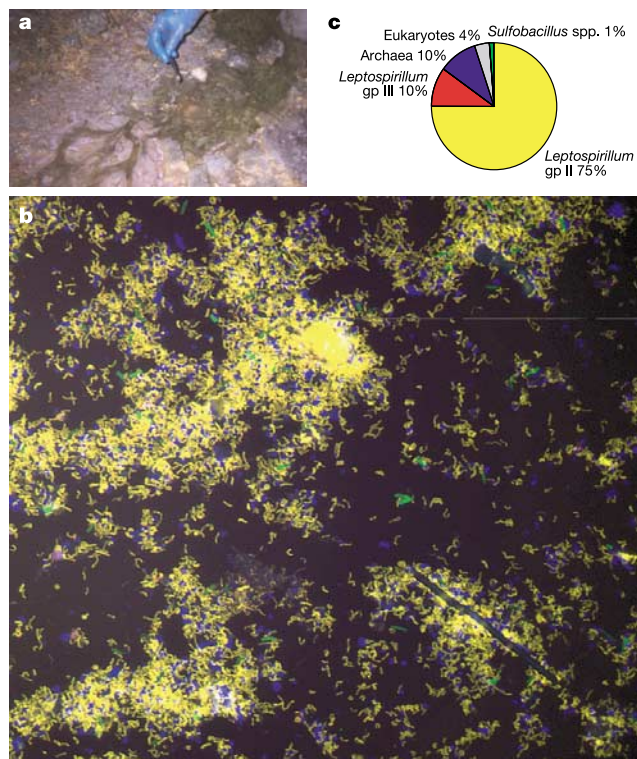


Figure 1 The pink biofilm. **a**, Photograph of the biofilm in the Richmond mine (hand included for scale). **b**, FISH image of **a**. Probes targeting bacteria (EUBmix; fluorescein isothiocyanate (green)) and archaea (ARC915; Cy5 (blue)) were used in combination with a probe targeting the *Leptospirillum* genus (LF655; Cy3 (red)). Overlap of red and green (yellow) indicates *Leptospirillum* cells and shows the dominance of *Leptospirillum*. **c**, Relative microbial abundances determined using quantitative FISH counts.

These results, and the agreement between the recovered and anticipated genome sizes, confirm that dividing scaffolds by G+C content, read depth and homology to the *fer1* genome is a valid means of sorting most genomes from this community data set. Methods for the analysis of genome signatures currently under development may be required for binning genome data from more complex communities^{24,25}.

Population structure and speciation

The biofilm sample contained approximately 10^8 *Leptospirillum* group II cells. Thus, the roughly 29,000 reads assembled into the *Leptospirillum* group II genome probably all came from different individuals. Despite this, there were no nucleotide polymorphisms in the 16S or 23S rRNA genes, or in their intergenic region. The average nucleotide polymorphism rate in the *Leptospirillum* group II genome is 0.08%; approximately one-third of the polymorphisms cause changes at the protein-coding level. The low incidence of nucleotide polymorphisms indicates that we have sequenced essentially a single strain from the community. Homogeneity within the *Leptospirillum* group II genome may reflect strong recent environmental selection for this genome type or be the result of a founder effect.

There are no nucleotide polymorphisms in the rRNA genes of *Ferroplasma* type II, despite an average polymorphism rate of about 2.2% (see Supplementary Information). Polymorphism-free regions typically a few hundred base pairs long and up to several kilobases in length occur one to tens of kilobases apart on the *Ferroplasma* type II scaffolds (see Supplementary Information). Although these homogenous regions may contain conserved genes, there is no consistency in the function of proteins encoded by them (for example, Fig. 2a; see also Supplementary Fig. S8).

In any given region, typically between one and three distinct patterns of nucleotide polymorphism were observed in the assembled *Ferroplasma* type II composite genome (Fig. 2b). By using sequence read end-pair information, these nucleotide polymorphism patterns could be connected across homogenous regions. We could identify points within individual reads and between end pairs where one distinct nucleotide polymorphism pattern transitioned into another (Fig. 2b; see also Supplementary Fig. S8). The most likely explanation for this, and for the larger homogeneous regions, is that the *Ferroplasma* type II strains have undergone homologous recombination. It is unlikely that the reads with pattern transitions represent variants that arose simply through accumulation of nucleotide polymorphisms, because this

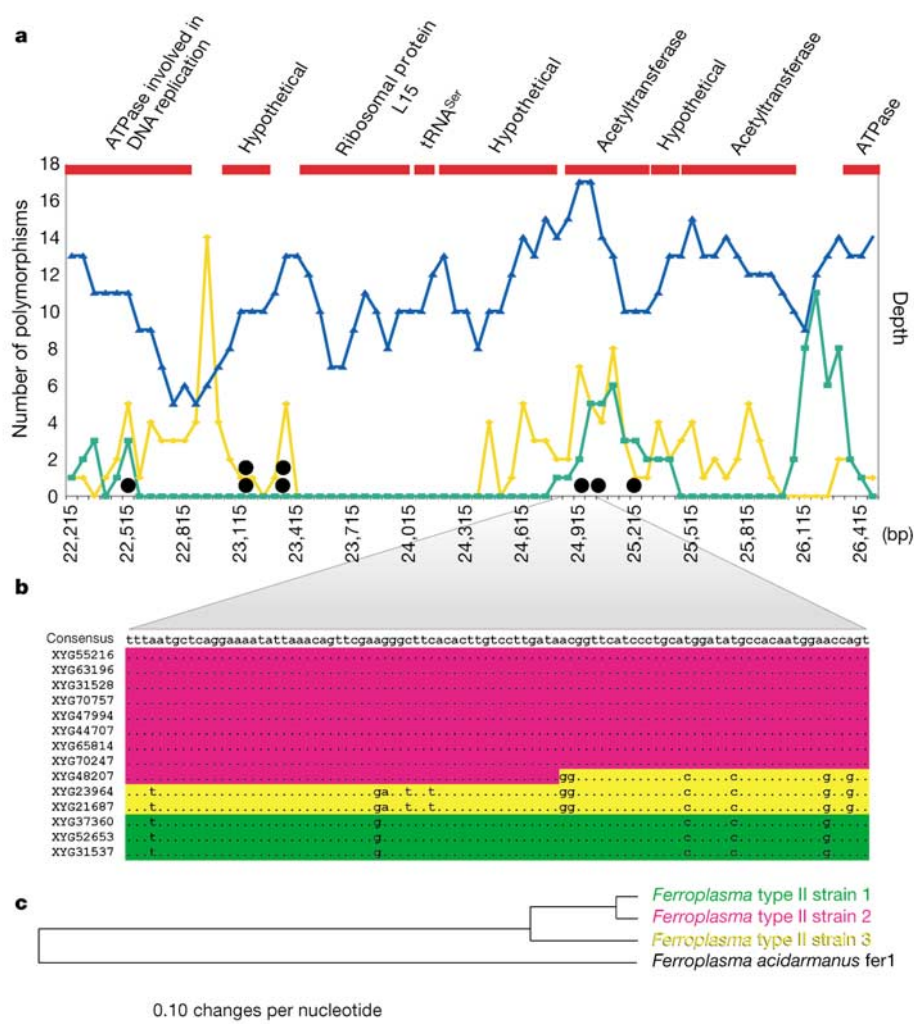


Figure 2 Segment of the *Ferroplasma* type II composite genome. **a**, A 4.2-kb region showing annotated open reading frames (ORFs) (red), average read depth (blue line), and the number of nucleotide polymorphisms in the 'green' and 'yellow' relative to the 'pink' strain (green and yellow lines) averaged over 60-bp windows. Black dots indicate

recombination sites. **b**, Alignment of individual reads (XYG) for a 96-bp region in **a**. Letters indicate nucleotide polymorphisms in the green and yellow strains relative to the pink strain. Note the recombinant sequence (XYG48207). **c**, Evolutionary distance tree inferred from the ancestral strain sequences in **a**.

would require precise selection acting on small degrees of heterogeneity at virtually every locus.

Assembled reads in regions of up to 10 kb chosen at random on *Ferroplasma* type II scaffolds generally contained about 20 transition points, interpreted to be recombination boundaries (see Supplementary Fig. S9). If the characteristics of these regions are typical, then switching of nucleotide polymorphism patterns occurs on average about every 5 kb, and the mosaic genomes of currently existing strains were constructed via at least 400 recombination events. It is impossible to obtain a detailed representation of the strain genome structure or relative abundances of individual types because we cannot directly link polymorphism patterns over long genomic regions with small insert library data.

At the time of sampling, the *Ferroplasma* type II species population seemed to be dominated by strains with mosaic genomes constructed by recombination of three closely related but distinct genome types (pink, green and yellow in Fig. 2b and Supplementary Fig. S9) that we infer correspond to three 'ancestral' strains (Fig. 3). Other ancestral types may have existed, and may have given rise to variants that are too rare to be detected by our analysis. Combinatorial variants such as these have been observed previously in populations of enteric bacteria from disparate locations²⁶ but they have not been documented in archaea or environmental samples from a single location.

Phylogenetic analysis of genome fragments reconstructed for the three ancestral strains reveals that they derived from a relatively recent common ancestor (Fig. 2c). On the basis of the overall strong internal consistency within the nucleotide polymorphism patterns, we infer that the most recent evolution of the *Ferroplasma* type II population has been dominated by homologous recombination. If the population were to undergo further recombination (a likely scenario), three distinct ancestral strains would still be identifiable at the local scale. Therefore, it is not possible to determine how many episodes of recombination have occurred in the population.

Sequences reconstructed for the three ancestral types (Fig. 2b) were used to calculate relative polymorphism frequencies (Fig. 2a). Most of the proteins encoded by the sequences were slightly different at the amino acid level. In one region (Supplementary Fig. S9) the nucleotide polymorphism rates were 0.6% for the green compared with pink (56% non-synonymous) and 3% for yellow compared with pink (44% non-synonymous). This suggests that recombination involving fragments carrying slightly different protein-coding sequences yields a very large number of genomic

combinatorial variants (Fig. 3) with subtly different metabolic characteristics. The existence of mosaic genome types has ecological and evolutionary significance because genome diversity due to extensive recombination would ensure availability of an optimized strain when the system is perturbed, conferring resilience to the species.

The frequency of recombination between organisms decreases exponentially as they become more divergent²⁷. Unlike the *Ferroplasma* type II strains, the *Ferroplasma* type I and II genomes have no anomalous regions of high sequence identity, indicating that they have not undergone recent recombination (at least on length scales smaller than the scaffold size). On the basis of the low recombination rate and the separation of these genomes from each other by assembly, we predict that *Ferroplasma* type I and II are physiologically distinct, and thus are separate species. Recombination and assembly may provide useful genome-based criteria to separate species from strains in cases where one or both organisms are uncultivated.

The combination of a 16S rRNA-based survey with comprehensive genomic sampling provides a snapshot of the population structure. We have not yet determined how stable the community structure is, or the factors responsible for the success of the observed strains. However, an important finding is the dominance of the biofilm by a handful of distinct genome types. A few organism types within a much larger pool of rare types is typical of lognormal abundance distributions in other natural communities^{28–30}. This may be attributed to a small number of niches within the AMD system at any one time, possibly because the ecosystem is relatively geochemically simple (for example, the dominant electron donors and acceptors are iron, sulphur and oxygen, and temperature and fluid composition cycle within a relatively narrow range over annual timescales)^{12,14}.

Pathways for genetic exchange

Although there is evidence for genetic exchange between *Ferroplasma* type II strains, it is unclear how recombination is achieved. There is also some evidence of ancient gene transfer between the Sulfolobales and members of the Thermoplasmatales, and between bacteria and archaea (data not shown). Transformation by uptake of naked DNA is unlikely, as DNA is not expected to have a significant residence time in the acid solution. There is no evidence for conjugation genes in *Ferroplasma* type I or II, and there is only limited evidence for transduction (some possible phage genes and integrases). We compared the sequences of the probable prophage genes in order to test whether the host range of phage in the AMD system is large enough to provide a mechanism for lateral gene transfer. Identical reverse transcriptases (LambdaSa1) occur in the *G-plasma* and *Ferroplasma* type II genomes (in very different genomic contexts), suggesting that a single phage type has recently targeted both lineages. Similarly, identical retron-type reverse transcriptases with identical adjacent transposases occur in otherwise different genomic contexts within the *Leptospirillum* group II and III genomes, indicating that a broad host range phage targets both of these groups.

Metabolic analysis

We recovered near-complete gene inventories for the five dominant members of the biofilm community. The data for *Leptospirillum* group II are particularly notable, as no genome of a *Nitrospira* phylum member had been sequenced previously. Here we focus on the metabolic pathways recovered in *Leptospirillum* group II and *Ferroplasma* type II (Fig. 4; see also annotation files in the Supplementary Information), and on the new insights into the ecological roles of individual members that have changed our understanding of how the community functions.

The acidophilic biofilms are self-sustaining communities that grow in the deep subsurface and receive no significant

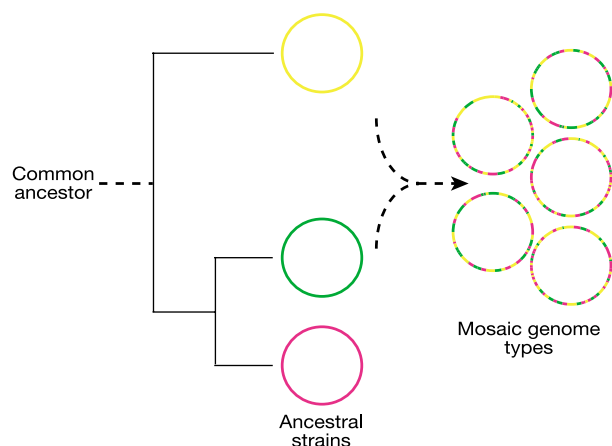
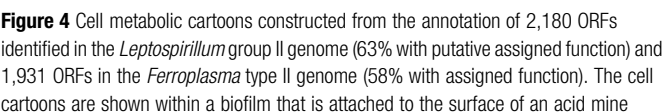


Figure 3 Schematic diagram illustrating a diversity of mosaic genome types within the *Ferroplasma* type II population that are inferred to have arisen by homologous recombination between three closely related ancestral genome types (pink, yellow and green).

contain formate hydrogenlyase complexes. These, in combination with carbon monoxide dehydrogenase, may be used for carbon fixation via the reductive acetyl coenzyme A (acetyl-CoA) pathway by some, or all, organisms. Given the large number of ABC-type sugar and amino acid transporters encoded in the *Ferroplasma* type



drainage stream (viewed in cross-section). Tight coupling between ferrous iron oxidation, pyrite dissolution and acid generation is indicated. Rubisco, ribulose 1,5-bisphosphate carboxylase-oxygenase. THF, tetrahydrofolate.

I and II and G-plasma genomes, these organisms may favour the heterotrophic lifestyle reported for other Thermoplasmatales³¹.

Nitrogen fixation is another process essential to the community. Given that *Leptospirillum ferrooxidans* (group I) possesses a complete nitrogen fixation pathway³² and because *Leptospirillum* group II dominates most biofilms in the AMD system, we suspected that *Leptospirillum* group II would be the primary nitrogen fixer. However, the *Leptospirillum* group II genome does not contain any genes for nitrogen fixation. We detected a complete nitrogen fixation operon with homology to that reported for a *L. ferrooxidans* strain³² in scaffolds assigned to *Leptospirillum* group III (for example, the *nifH* gene is 84% identical to that of *L. ferrooxidans*). Manual linking of this scaffold to adjacent scaffolds with low coverage supports the assignment of the *nif* operon to *Leptospirillum* group III. In order to confirm this assignment we used primers to amplify *nifH* genes. *nifH* was recovered from the community DNA but not from a mixed culture of *Leptospirillum* group II strains obtained from the site (Tyson *et al.*, unpublished data).

There is no evidence to suggest that *Ferroplasma* type I or II fix nitrogen. It is likely that *Ferroplasma* species obtain nitrogen fixed by other organisms from solution using the numerous amino acid transporters and ammonia permeases (Fig. 4). As the community-essential role of nitrogen fixation is assigned to a relatively low-abundance organism type, *Leptospirillum* group III may be a key-stone species in this ecosystem.

The use of ferrous iron oxidation as an energy source directly links microbial and geochemical processes in AMD ecosystems. Iron oxidation has been well studied only in *Acidithiobacillus ferrooxidans*^{33–35}, despite *Leptospirillum* and *Ferroplasma* being the dominant iron oxidizers in many AMD systems¹⁴. We have reconstructed a putative electron transport chain for *Leptospirillum* group II (Fig. 4). The genome does not encode the blue copper protein³⁴ or putative direct ferrous iron oxidase^{35,36}, which are present in *A. ferrooxidans*; however, a number of novel cytochromes were detected (Ram *et al.*, unpublished data).

All aerobic iron-oxidizing organisms in the AMD system face the challenge of generating energy in microaerophilic solutions. *Leptospirillum* group II and III genomes both encode genes indicative of cytochrome *cbb*₃-type haeme-copper oxidases and cytochrome *bd*-type quinol oxidases, both of which typically have high affinity for oxygen^{37,38}. These genes may protect the *Leptospirillum* group III oxygen-sensitive nitrogenase complex during nitrogen fixation, as proposed in other systems³⁹.

The electron transport chain of *Ferroplasma* type II differs significantly from that in *Leptospirillum* group II (Fig. 4). It contains putative haeme-copper terminal oxidases, cytochrome *b* and associated Rieske iron-sulphur proteins, and a blue copper protein, all of which may be assembled into a terminal oxidase supercomplex similar to SoxM in *Sulfolobus acidocaldarius*⁴⁰. Interestingly, the blue copper protein shares sequence characteristics with both *A. ferrooxidans* rusticyanins and *Sulfolobus* sulphocyanins, thus it may be a component of the electron transport chain during iron oxidation as well as heterotrophic growth. As with *Sulfolobus*, *Ferroplasma* type I and II and G-plasma do not contain coding regions for cytochrome *c*.

The robust macroscopic biofilms provide attachment to the surroundings and allow them to float at the air–water interface, simultaneously optimizing access to O₂, CO₂, N₂ and dissolved ferrous iron. It is unknown which organisms are responsible for polymer production. The *Leptospirillum* group II genome contains an operon of putative cellulose synthase genes. Cellulose has been shown to prevent desiccation and to enable biofilms to float. Numerous glycosyltransferases and polysaccharide export proteins also suggest a role for *Leptospirillum* group II in biofilm formation. In contrast, there is little evidence for genes for extracellular polymer production in the *Ferroplasma* type II genome.

After perturbation, the first colonizers are probably those that are

motile. Both *Leptospirillum* group II and III possess a chemotaxis system and flagellar operon, which may be used to respond to ferrous iron⁴¹ and oxygen gradients. There is no evidence for flagellar components or of other genes required for motility in any of the archaeal genomes. This explains the small number of proteins in *Ferroplasma* type II that are assigned to the cell motility and secretion (see Supplementary Fig. S10).

Perhaps the most important challenges facing members of biofilm communities growing in extremely acidic, metal-rich solutions are maintaining a near-neutral cytoplasm and resistance to toxic metals. *Ferroplasma* type II has genes for production of isoprenoid-based lipids, which we presume are largely tetraether linked⁴², and thus highly proton impermeable. *Leptospirillum* group II seems to dedicate a comparatively large number (see Supplementary Fig. S10) and variety of genes for cell membrane biosynthesis, suggesting a complex cell wall structure. In addition, there are a variety of proton efflux systems ((H⁺)ATPases, antiporters and symporters). Most community members possess genes for resistance to copper, cobalt, arsenite, mercury, zinc, silver and cadmium (for example, *merA*, *arsB*, *arsR*).

An intriguing finding is the presence of genes for proteins related to DNA photolyases in the *Leptospirillum* group II (SplB and PhrB), III (PhrB) and G-plasma genomes (PhrB). The two putative photolyases in G-plasma are most closely related to photolyases of bacterial origin. These typically light-activated proteins for the repair of ultraviolet-damaged DNA may be present because relatively recent ancestors of these strains populated surface environments.

Concluding remarks

The culture-independent recovery of two near-complete microbial genomes and partial recovery of three other genomes from an environmental sample is an advance in the study of natural microbial communities. Genome reconstruction was facilitated by the fact that the community was dominated by populations of a few genomically distinct species. The effectiveness of this approach in other environments will be limited by high species richness, heterogeneities in the abundance of community members, as well as by extensive genome rearrangements. However, even in more complex environments, it should be possible to extend the random shotgun sequencing approach to recover the genomes of uncultivated strains and species. These data can then be used to explore the nature of the community metabolic network, to find conditions for cultivating previously uncultivated organisms, to monitor community structure over time, and to construct DNA microarrays to monitor global community gene expression patterns. □

Methods

The biofilm sample chosen for genomic analysis was characterized using FISH to quantify the abundance of the dominant prokaryotic members. Environmental DNA extracted using an agarose plug method was used for construction of 16S rRNA gene and small insert (approximately 3 kb) pUC18 libraries. Double-ended sequencing reactions were carried out using PE BigDye terminator chemistry (Perkin Elmer) and resolved using an ABI PRISM 3730 (Applied Biosystems) capillary DNA sequencer. Vector and quality trimming of shotgun data was performed to yield 103,462 reads totalling 76.2 Mb (average trimmed read length of 737 bp). After assembly and assignment of scaffolds to organism types, we identified the protein and RNA genes using the Fgenesb_annotator pipeline (Softberry Inc.). The gene prediction algorithm is based on Markov chain models of coding regions and translation and termination sites. We used the tRNAscan-SE package for prediction of tRNA genes. All reads were aligned to the scaffolds using BLASTN for analysis of nucleotide polymorphisms. Details for all methods are provided in the Supplementary Information.

Received 24 November 2003; accepted 19 January 2004; doi:10.1038/nature02340.

Published online 1 February 2004.

- Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
- Makarova, K. S. & Koonin, E. V. Comparative genomics of archaea: how much have we learned in six years, and what's next? *Genome Biol.* **4**, 115.1–115.16 (2003).
- Koonin, E. V. & Mushegian, A. R. Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr. Opin. Genet. Dev.* **6**, 757–762 (1996).
- Amann, R. L., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and *in-situ* detection of individual microbial-cells without cultivation. *Microbiol. Rev.* **59**, 143–169 (1995).

5. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
6. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**, reviews0003.1–0003.8. (2002).
7. Beja, O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2**, 516–529 (2000).
8. Beja, O. *et al.* Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* **68**, 335–345 (2002).
9. Rondon, M. R. *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).
10. Torsvik, V., Ovreas, L. & Thingstad, T. F. Prokaryotic diversity—magnitude, dynamics, and controlling factors. *Science* **296**, 1064–1066 (2002).
11. Singer, P. C. & Stumm, W. Acidic mine drainage rate-determining step. *Science* **167**, 1121–1127 (1970).
12. Edwards, K. J., Gihring, T. M. & Banfield, J. F. Seasonal variations in microbial populations and environmental conditions in an extreme acid mine drainage environment. *Appl. Environ. Microbiol.* **65**, 3627–3632 (1999).
13. Bond, P. L., Smriga, S. P. & Banfield, J. F. Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. *Appl. Environ. Microbiol.* **66**, 3842–3849 (2000).
14. Bond, P. L., Druschel, G. K. & Banfield, J. F. Comparison of acid mine drainage microbial communities in physically and geochemically distinct ecosystems. *Appl. Environ. Microbiol.* **66**, 4962–4971 (2000).
15. Baker, B. J. & Banfield, J. F. Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol.* **44**, 139–152 (2003).
16. Edwards, K. J. *et al.* Geochemical and biological aspects of sulfide mineral dissolution: lessons from Iron Mountain, California. *Chem. Geol.* **169**, 383–397 (2000).
17. Silverman, M. P. & Ehrlich, H. L. Microbial formation and degradation of minerals. *Adv. Appl. Microbiol.* **6**, 153–206 (1964).
18. Bond, P. L. & Banfield, J. F. Design and performance of rRNA targeted oligonucleotide probes for *in situ* detection and phylogenetic identification of microorganisms inhabiting acid mine drainage environments. *Microb. Ecol.* **41**, 149–161 (2001).
19. Coram, N. J. & Rawlings, D. E. Molecular relationship between two groups of the genus *Leptospirillum* and the finding that *Leptospirillum ferriphilum* sp nov dominates South African commercial biooxidation tanks that operate at 40 °C. *Appl. Environ. Microbiol.* **68**, 838–845 (2002).
20. Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T. & Salzberg, S. L. The value of complete microbial genome sequencing (you get what you pay for). *J. Bacteriol.* **184**, 6403–6405 (2002).
21. Branscomb, E. & Predki, P. On the high value of low standards. *J. Bacteriol.* **184**, 6406–6409 (2002).
22. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
23. Amils, R., Irazabal, N., Moreira, D., Abad, J. P. & Marin, I. Genomic organization analysis of acidophilic chemolithotrophic bacteria using pulse field gel electrophoretic techniques. *Biochimie* **80**, 911–921 (1998).
24. Sandberg, R. *et al.* Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res.* **11**, 1404–1409 (2001).
25. Abe, T. *et al.* Informatics for unveiling hidden genome signatures. *Genome Res.* **13**, 693–702 (2003).
26. Spratt, B. G., Hanage, W. P. & Feil, E. J. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr. Opin. Microbiol.* **4**, 602–606 (2001).
27. Vulic, M., Dionisio, F., Taddei, F. & Radman, M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl Acad. Sci. USA* **94**, 9763–9767 (1997).
28. Magurran, A. E. & Henderson, P. A. Explaining the excess of rare species in natural species abundance distributions. *Nature* **422**, 714–716 (2003).
29. McGill, B. J. A test of the unified neutral theory of biodiversity. *Nature* **422**, 881–885 (2003).
30. Dewdney, A. K. A dynamical model of communities and a new species-abundance distribution. *Biol. Bull.* **198**, 152–165 (2000).
31. Ruepp, A. *et al.* The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407**, 508–513 (2000).
32. Parro, V. & Moreno-Paz, M. Gene function analysis in environmental isolates: The *nif* regulon of the strict iron oxidizing bacterium *Leptospirillum ferrooxidans*. *Proc. Natl Acad. Sci. USA* **100**, 7883–7888 (2003).
33. Blake, R., Shute, E. A., Waskovsky, J. & Harrison, A. P. Respiratory components in acidophilic bacteria that respire on iron. *Geomicrobiol. J.* **10**, 173–192 (1992).
34. Blake, R. & Shute, E. A. Respiratory enzymes of *Thiobacillus ferrooxidans*—kinetic-properties of an acid-stable iron-rusticyanin oxidoreductase. *Biochemistry* **33**, 9220–9228 (1994).
35. Yamanaka, T. & Fukumori, Y. Molecular aspects of the electron transfer system which participates in the oxidation of ferrous ion by *Thiobacillus ferrooxidans*. *FEMS Microbiol. Rev.* **17**, 401–413 (1995).
36. Appia-Ayme, C., Guilian, N., Ratouchniak, J. & Bonnefoy, V. Characterization of an operon encoding two *c*-type cytochromes, an *aa(3)*-type cytochrome oxidase, and rusticyanin in *Thiobacillus ferrooxidans* ATCC 33020. *Appl. Environ. Microbiol.* **65**, 4781–4787 (1999).
37. Preisig, O., Zufferey, R. & Hennecke, H. The *Bradyrhizobium japonicum* fixGHIS genes are required for the formation of the high-affinity *cbb(3)*-type cytochrome oxidase. *Arch. Microbiol.* **165**, 297–305 (1996).
38. Pitcher, R. S., Brittain, T. & Watmough, N. J. Cytochrome *cbb(3)* oxidase and bacterial microaerobic metabolism. *Biochem. Soc. Trans.* **30**, 653–658 (2002).
39. Poole, R. K. & Hill, S. Respiratory protection of nitrogenase activity in *Azotobacter vinelandii*—roles of the terminal oxidases. *Biosci. Rep.* **17**, 303–317 (1997).
40. Komorowski, L., Verheyen, W. & Schafer, G. The archaeal respiratory supercomplex SoxM from *S. acidocaldarius* combines features of quinole and cytochrome *c* oxidases. *Biol. Chem.* **383**, 1791–1799 (2002).
41. Acuna, J., Rojas, J., Amaro, A. M., Toledo, H. & Jerez, C. A. Chemotaxis of *Leptospirillum ferrooxidans* and other acidophilic chemolithotrophs—comparison with the *Escherichia coli* chemosensory system. *FEMS Microbiol. Lett.* **96**, 37–42 (1992).
42. Macalady, J. L. *et al.* Tetraether-linked membrane monolayers in *Ferroplasma* spp.: a key to survival in acid. *Extremophiles* (submitted).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This research was funded by the US Department of Energy Microbial Genomics Program and the National Science Foundation Biocomplexity Program. We would like to thank M. Power, W. Getz, R. Blake, J. Handlesman, B. Baker, I. Lo, J. Flanagan, D. Dodds and R. Carver for their contributions to this work. We also thank C. Detter and members of his laboratory at JGI for help with library construction, and T. Arman (Iron Mountain Mines) and R. Sugarek (EPA) for access to the Richmond mine.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.F.B. (jill@eps.berkeley.edu). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession code AADL00000000. The version described in this paper is the first version, AADL01000000.

Mice cloned from olfactory sensory neurons

Kevin Eggan^{1*†}, Kristin Baldwin^{2*}, Michael Tackett¹, Joseph Osborne^{2†}, Joseph Gogos², Andrew Chess¹, Richard Axel² & Rudolf Jaenisch¹

¹Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

²Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, College of Physicians and Surgeons, Columbia University, 701 West 168th Street, New York, New York 10032, USA

* These authors contributed equally to this work

† Present addresses: Department of Physiology and Cellular Biophysics, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, New York, New York 10032, USA (J.O.); Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA (K.E.)

Cloning by nuclear transplantation has been successfully carried out in various mammals, including mice. Until now mice have not been cloned from post-mitotic cells such as neurons. Here, we have generated fertile mouse clones derived by transferring the nuclei of post-mitotic, olfactory sensory neurons into oocytes. These results indicate that the genome of a post-mitotic, terminally differentiated neuron can re-enter the cell cycle and be reprogrammed to a state of totipotency after nuclear transfer. Moreover, the pattern of odorant receptor gene expression and the organization of odorant receptor genes in cloned mice was indistinguishable from wild-type animals, indicating that irreversible changes to the DNA of olfactory neurons do not accompany receptor gene choice.

The chromosome is a dynamic structure that undergoes complex changes that underlie the development and differentiated function of an organism. Alterations in chromosome structure include variation in the complement of regulatory proteins, covalent modifications in chromatin proteins or DNA, and in rare instances, DNA rearrangements¹. The extent to which such chromosomal changes are reversible can be discerned by cloning experiments involving nuclear transfer. Thus far, cloning experiments using nuclei from post-mitotic cells that have irreversibly exited the cell cycle as part of their programme of differentiation have not generated viable embryos or mice^{2,3}. These observations have led to the suggestion that post-mitotic cells might be refractory to epigenetic reprogramming or alternatively might have acquired changes in their DNA that could limit their developmental potential^{4,5}.

Directed DNA rearrangements are rarely observed as part of a normal differentiation programme and, in vertebrates, they have been described only for the generation of the diverse repertoire of antibodies and T-cell receptors^{6–8}. Several observations have suggested that post-mitotic neurons may also use irreversible DNA alterations to generate diversity. First, a population of neurons undergoes apoptosis in mice bearing mutations in the double-strand-break DNA-repair enzymes, which are also required for DNA rearrangements in lymphocytes^{9,10}. Second, cortical neurons exhibit a higher incidence of aneuploidy than other cell types, although the functional significance of these changes is unknown^{11,12}. These observations and the inability to clone mice from neuronal nuclei have led to models in which the DNA of post-mitotic neurons might undergo rearrangements to supply additional genetic diversity that may enhance neural function^{4,5,13–15}.

One particularly clear example of neuronal diversity is provided by the olfactory sensory epithelium. In the mouse, each of the 2,000,000 cells in the olfactory epithelium expresses only one of about 1,500 odorant receptor genes, such that the functional identity of a neuron is defined by the nature of the receptor it expresses¹⁶. Thus, the sensory epithelium consists of at least 1,500 neuronal types. The pattern of receptor expression is apparently random within one of four zones in the epithelium, suggesting that

the choice of receptor gene may be stochastic^{17,18}. One mechanism to permit the stochastic choice of a single receptor could involve DNA rearrangements^{19,20}.

Here we report the generation of fertile mouse clones by transferring the nuclei of post-mitotic olfactory neurons into enucleated oocytes. Thus, a post-mitotic nucleus can re-enter the cell cycle and be reprogrammed to totipotency. The DNA of mice derived from sensory neurons reveals no evidence for rearrangements of the expressed olfactory receptor gene. The pattern of receptor expression in these mice was indistinguishable from that of wild-type animals, indicating that irreversible changes in DNA do not accompany olfactory receptor gene choice.

Genetic marking of olfactory sensory neurons

Less than 1% of nuclear transfers result in the production of an embryonic stem (ES) cell line or live animal¹. It is therefore difficult to identify with certainty the origin of the donor nucleus that contributed to the generation of a particular cloned animal². This problem is particularly apparent in the olfactory epithelium, which contains mature sensory neurons intermingled with stem cells, neural progenitors and support cells. Therefore, we generated mice in which the endogenous olfactory marker protein (OMP) promoter drives simultaneous expression of OMP and Cre recombinase by inserting an internal ribosome entry site (IRES)-Cre cassette 3' of the OMP stop codon. OMP is expressed only in mature olfactory sensory neurons (OSNs). When OMP-IRES-Cre mice are crossed to a reporter mouse strain (Z/EG), Cre expression catalyses the excision of a transcriptional stop sequence and results in green fluorescent protein (GFP) expression solely in mature OSNs (Fig. 1a and Methods). Transfer of marked OSN nuclei into oocytes should generate embryos in which every cell expresses GFP, as the Z/EG reporter uses the ubiquitous actin promoter.

In mice bearing the OMP-IRES-Cre allele and the Z/EG reporter gene, GFP expression was observed only in the most mature OSNs. Sections through the entire olfactory epithelium of several adult mice revealed GFP expression in the regions that contain mature OSNs (Fig. 1b–f). Cell counts revealed no GFP⁺ cells in either the basal cell layers, where immature progenitors and stem cells reside, or in the apical support cell layer (0 of 835 GFP⁺ cells)^{21,22}. Double

immunostaining for GFP and Cre protein revealed that all GFP⁺ cells with visible nuclei also expressed Cre recombinase (599 of 599) (Fig. 1c). In addition, none of the MASH-1⁺ basal cell precursors of OSNs expressed GFP (0 of 350 MASH-1⁺ cells) (Fig. 1d). These results showed that GFP expression was restricted to mature sensory neurons.

We confirmed that the GFP⁺ mature OSNs were post-mitotic by injecting mice with 5-bromodeoxyuridine (BrdU) and staining the olfactory epithelium with anti-BrdU and GFP antibodies (Fig. 1e). BrdU incorporation was observed in the more basal GFP-negative layers as well as in rare support cells in the apical layers. Examination of more than 3,000 GFP⁺ cells never revealed a GFP⁺ BrdU⁺ cell. In a separate experiment, labelling with an antibody specific to Ki67, a protein restricted to the nuclei of dividing cells, revealed no staining coincident with GFP⁺ (Fig. 1f). These data show that GFP⁺ cells in the olfactory epithelium of donor animals are mature post-mitotic OSNs.

Cloning mice from neuronal nuclei

We asked whether the nucleus of a post-mitotic OSN could re-enter the cell cycle and direct preimplantation development. We dissociated the olfactory epithelium and picked GFP-expressing OSNs for nuclear transfer into enucleated oocytes (Fig. 2a)^{25–29}. Of the 352 embryos generated, 48 (14%) developed to the blastocyst stage (Table 1). All blastocysts expressed GFP, demonstrating that they were derived from the mature OSN donor nuclei (Fig. 2b). We confirmed that Cre expression in the OSNs caused constitutive GFP expression after nuclear transfer by generating blastocysts with GFP-negative cells from the donor mice. As expected, none of these blastocysts expressed GFP, demonstrating that Cre is not aberrantly expressed as a result of nuclear transfer or subsequent cloning procedures.

The 48 GFP⁺ blastocysts were used to generate ES cell lines and three gave rise to colonies that resembled ES cells (OSN1–3). All three lines expressed GFP, and Southern blotting confirmed the predicted genomic rearrangement at the Z/EG locus (Fig. 2c and data not shown). When the OSN2 and OSN3 ES cells were injected into diploid blastocysts, they contributed extensively to all tissues of

the resulting chimaeras, including the germ line (Fig. 2i, j and data not shown).

We injected these cloned ES cells into tetraploid blastocysts to perform the most stringent test of their developmental potency (Table 2). Tetraploid embryo complementation generates an ‘ES-fetus’ in which all embryonic lineages are derived from the injected ES cells, whereas the extraembryonic lineages develop from the host blastocyst^{29,30}. Both OSN2 and OSN3 ES cells gave rise to embryos of embryonic day (E)19.5 that expressed GFP ubiquitously (Table 2 and Fig. 2d). Mice derived from OSN3 survived to adulthood, were fertile and were overtly normal. Histological examination of serial sections through their brains revealed no obvious abnormalities (data not shown). These results indicate that the genome of a post-mitotic, terminally differentiated neuron can be reprogrammed after nuclear transfer and direct the development of all embryonic lineages.

Odorant receptor expression in cloned mice

We examined the patterns of receptor gene expression in the olfactory epithelium of the cloned mice generated by tetraploid complementation. If irreversible genetic rearrangements are required for receptor choice we might expect an altered profile of receptor expression in animals cloned from an OSN. In the simplest model, a rearrangement involving one receptor gene might persist in all neurons such that all neurons will express the same receptor. This scenario was observed in mice derived from a B-cell nucleus, in which all B cells expressed the same immunoglobulin gene²⁸. We therefore asked whether sensory neurons from cloned mice express a single receptor or a repertoire of receptor genes. We performed *in situ* hybridization on the olfactory epithelium of mice derived from OSN3 ES cells using probes specific for seven different odorant receptors (Fig. 2e–h and data not shown). The pattern of expression of all seven receptors was indistinguishable from control mice, excluding a simple model of gene rearrangement that would result in the expression of a single receptor gene in all OSNs.

As a neuron expresses a receptor from only one of the two alleles, it remained possible that a rearranged receptor gene expressed in the donor nucleus would be expressed in only half of the sensory

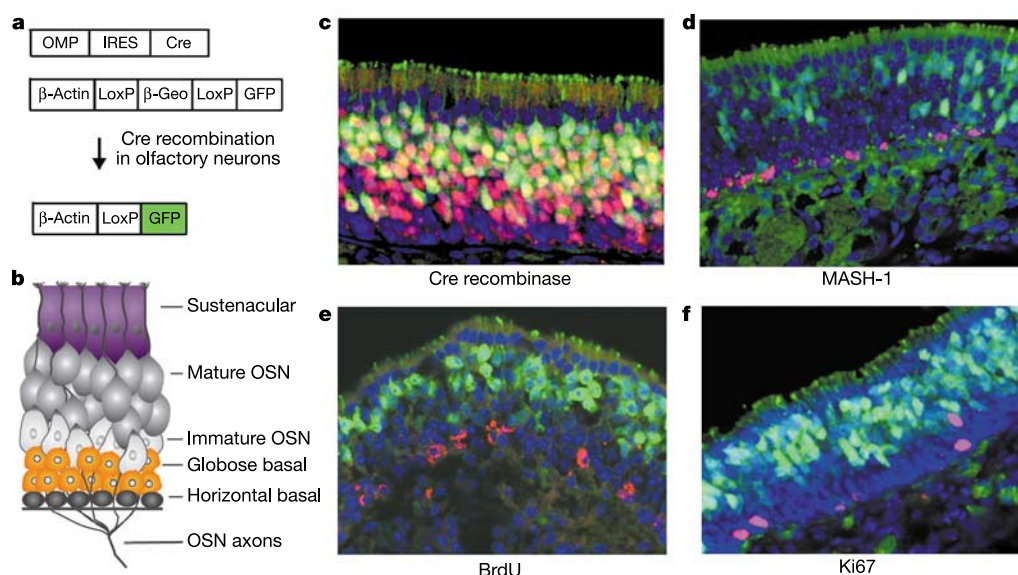


Figure 1 Genetically marking post-mitotic OSNs. **a**, Donor animals carry one OMP-IRES-Cre allele and one copy of the Z/EG Cre reporter transgene. Cre expression in OSNs catalyzes excision of the β -geo/stop cassette, resulting in selective GFP expression and genetic marking of mature OSNs. Thus, GFP⁺ neurons could be selectively chosen for nuclear transfer. **b**, Schematic diagram of the approximate laminar distribution of cell

types in the mouse olfactory epithelium. **c–f**, Three-colour immunofluorescence staining. Blue, nuclear marker TOTO-3; green, GFP protein fluorescence (**c**, **d**) or antibody to GFP protein (**e**, **f**); red, antibodies to Cre recombinase (**c**), progenitor-specific marker MASH-1 (**d**), markers of dividing cells, BrdU (**e**) and Ki67 (**f**).

neurons. We therefore analysed the repertoire of receptors expressed by polymerase chain reaction with reverse transcription (RT-PCR) to determine whether a single receptor transcript was enriched in the epithelium of OSN-derived mice. RNA from the olfactory epithelium of cloned animals was used in RT-PCR reactions with degenerate primers that recognize conserved motifs present in the majority of odorant receptors. Forty-four PCR products were cloned and restriction digest analysis indicated that they encoded 38 different receptors. Sequence analysis of 20 of these PCR clones revealed that they encoded 20 different receptors, which were located in seven clusters on six different chromosomes. This suggests that a single odorant receptor did not predominate in the olfactory epithelium of mice derived from OSN nuclei.

One additional assay for the diversity of odorant receptor expression is based on the observation that neurons expressing a given receptor, although randomly distributed within a zone of the epithelium, converge on two spatially invariant loci or glomeruli in the olfactory bulb^{18,31}. If half of the OSNs from cloned mice

expressed the same receptor, then their axonal projections should preferentially innervate a small set of glomeruli. Analysis of the olfactory bulb of chimaeric mice produced with OSN2 ES cells revealed that the OSN2-derived GFP⁺ cells innervated most, if not all glomeruli, with no glomerulus receiving predominant input (Fig. 2i, j). These results show that the sensory neurons of mice cloned from an OSN that had expressed a single receptor can express a large repertoire of odorant receptor genes.

Mice cloned from neurons expressing the P2 receptor

The previous experiments suggest that irreversible rearrangements are not required for receptor gene expression. Because these experiments did not allow the prospective identification of the receptor gene expressed by the donor OSN nucleus, we were unable to examine its pattern of expression or its DNA sequence in the cloned mice. Neurons that express the P2 odorant receptor were marked by introducing an IRES directing the translation of GFP into the 3' untranslated region of the P2 gene (P2-IRES-GFP)³².

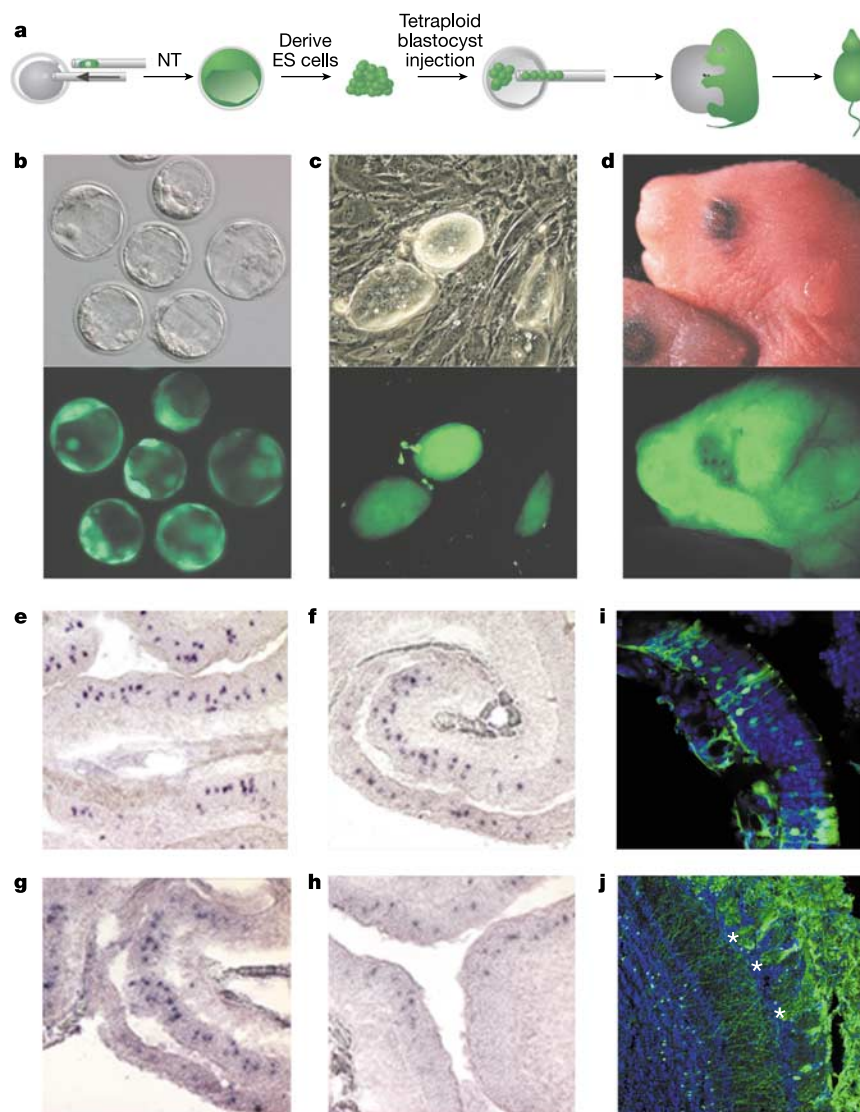


Figure 2 Mice derived from OSN nuclei. **a**, Strategy to generate cloned ES cell lines and OSN-derived mice by tetraploid complementation with cloned ES cells. **b, c**, Bright-field and fluorescence images of nuclear transfer blastocysts (**b**) and OSN3 ES cells (**c**) produced from OSN nuclei. **d**, A P0.5 mouse produced by tetraploid complementation with OSN3 ES cells (top) and a P0.5 C57/B6 control (bottom). **e-h**, *In situ* hybridization on

the olfactory epithelium of mice wholly derived from OSN3 ES cells with probes for odorant receptors P2 (**e**), M50 (**f**), I7 (**g**) and M71 (**h**). **i, j**, Contribution of GFP⁺ OSN2 ES cells to the olfactory epithelium (**i**) and olfactory bulb (**j**) of chimaeras generated by injecting OSN2 ES cells into diploid blastocysts. Several representative glomeruli are demarcated with asterisks.

Table 1 Production of ES cell lines and mice by nuclear transfer

Donor cells	Oocytes surviving (% injected)	Oocytes activated (% surviving)	Two-cell embryos to oviduct (% oocytes activated)	Blastocysts (% oocytes activated)	ES cell lines (% oocytes activated)	Alive at term (% into oviduct)
OSNs	508 (88)	352 (69)	—	48 (14)	3 (1)	—
P2 OSNs	261 (95)	141 (54)	—	18 (13)	2 (1)	—
P2SN1 ES	50 (95)	38 (76)	14 (37)	—	—	1 (7)
P2SN2 ES	47 (90)	37 (79)	18 (49)	—	—	1 (6)

Table 2 Mice produced from sensory neurons by tetraploid embryo complementation

ES cell line	Receptor expressed	Tetraploid blastocysts injected	Mice alive at term (% injected)	Breathing normally (% injected)	Cross-fostered	Survival to maturity (% fostered)
OSN1	?	137	0	0	0	0
OSN2	?	160	1 (1)	0	0	0
OSN3	?	273	49 (18)	29 (11)	21	17 (81)
P2SN1	P2	89	16 (18)	16 (100)	13	12 (92)
P2SN2	P2	151	31 (21)	28 (90)	28	22 (78)

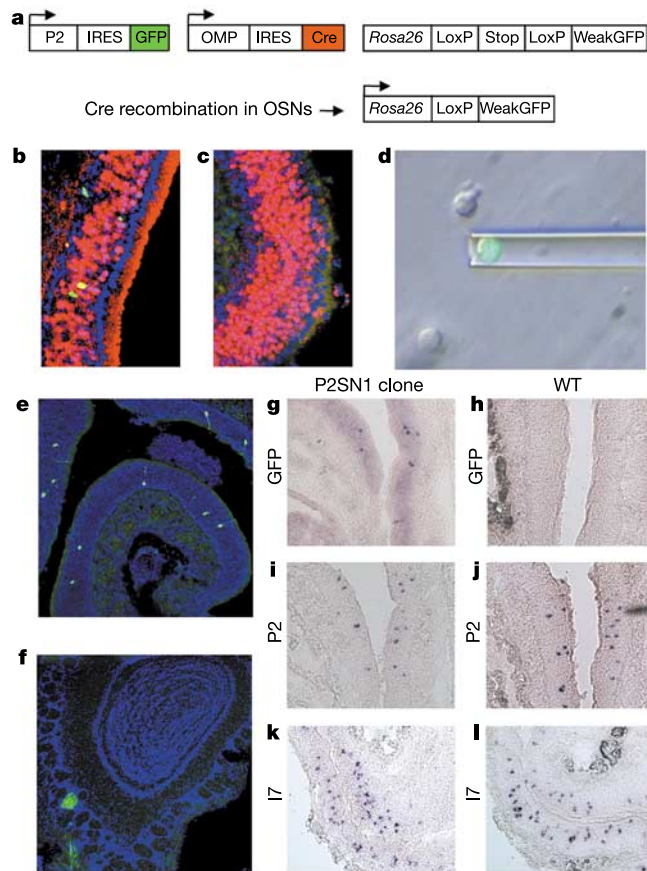


Figure 3 Mice produced from OSNs expressing the P2 odorant receptor. **a**, Strategy to label P2 sensory neurons with GFP and genetically mark mature OSNs. **b, c**, Immunofluorescence staining with nuclear marker TOTO-3 (blue), GFP fluorescence (green) and antibodies to Cre recombinase (red) of the olfactory epithelium of donor animals heterozygous for the P2-IRES-GFP, OMP-IRES-Cre and *Rosa26*-LoxP-Stop-LoxP-weak GFP alleles (**b**) or of control animals heterozygous for the OMP-IRES-Cre and *Rosa26*-LoxP-Stop-LoxP-weak GFP alleles (**c**). **d**, Bright-field and fluorescence merged image of picking a P2-IRES-GFP⁺ neuron from the dissociated olfactory epithelium of a donor for nuclear transfer. **e, f**, GFP fluorescence (green) and TOTO-3⁺ nuclei (blue) visualized in sections of the olfactory epithelium (**e**) and olfactory bulb (**f**) of an animal generated by tetraploid complementation with P2SN1 ES cells. **g–l**, *In situ* hybridization on olfactory epithelium sections derived entirely from P2SN2 ES cells (**g, i, k**) and a wild-type control (**h, j, l**) with probes for GFP (**g, h**), P2 receptor (**i, j**) and I7 receptor (**k, l**).

These mice were crossed with strains carrying both the OMP-IRES-Cre alteration and a reporter allele in which the *Rosa26* promoter is separated from a weak GFP gene by a LoxP-flanked transcriptional terminator³³ (Fig. 3a). Mice carrying the P2-IRES-GFP, OMP-IRES-Cre and *Rosa26*-LoxP-Stop-LoxP-weak GFP reporter genes exhibited intense GFP fluorescence in the approximately 0.1% of the OSNs that expressed the P2-IRES-GFP allele, whereas all other neurons appeared dark (Fig. 3b). *Rosa26*-driven GFP was present in mature OSNs, but the GFP signal was too weak to be detected by direct fluorescence. No green cells were observed in animals that contained only the OMP-IRES-Cre and *Rosa26*-LoxP-Stop-LoxP-weak GFP genes (Fig. 3c). Evidence for Cre-mediated excision of the transcription terminator and the resulting expression of weak GFP in mature OSNs was detected by amplifying the signal with GFP antibodies (data not shown).

Fluorescent cells expressing the P2-IRES-GFP allele were picked and their nuclei were injected into enucleated oocytes (Fig. 3d). Eighteen blastocysts developed from 141 reconstructed oocytes generating two ES cell lines: P2SN1 and P2SN2 (Table 1). PCR analysis revealed Cre-mediated excision at the *Rosa26* locus in both cell lines, indicating their origin from cloned nuclei of post-mitotic OSNs (Supplementary Fig. 1). Tetraploid embryos injected with P2SN1 and P2SN2 cells resulted in the birth of multiple viable pups that survived to adulthood (Table 2). These mice exhibited no gross anatomic or behavioural abnormalities and were fertile. PCR analyses of non-neuronal tissues revealed recombination at the *Rosa26* reporter allele, indicating that the cloned mice were derived from mature OMP⁺ OSNs (data not shown). In addition, we detected the weak *Rosa26*-driven GFP expression by *in situ* hybridization in both the olfactory epithelium and non-neuronal tissues, demonstrating genetic activation of the reporter (Fig. 3g, h and data not shown).

The expression pattern of P2-IRES-GFP in clones was indistinguishable from that of control donor animals. Approximately 0.1% of the neurons expressed GFP at a high level within zone II of the epithelium (Fig. 3e). The GFP-expressing neurons projected axons to one medial and one lateral glomerulus in the olfactory bulb (Fig. 3f). Moreover, the P2-IRES-GFP allele was expressed at a frequency approximately equal to that of the unmodified P2 allele. *In situ* hybridization was performed on sections through the olfactory epithelium with RNA probes specific to GFP- and P2-coding sequences (Fig. 3g–j). The number of GFP-expressing cells in one section was roughly 50% of the number of P2-expressing cells found in neighbouring sections ($45 \pm 14\%$, s.d.), revealing no preference for the expression of the P2-IRES-GFP allele. Furthermore, *in situ* hybridization with probes specific to six additional

olfactory receptors revealed similar expression patterns in control and cloned mice (Fig. 3k, l and data not shown).

We also performed Southern blotting, PCR and genome sequencing to examine the organization of the P2-IRES-GFP allele in the chromosome of cloned mice in an effort to detect potential DNA rearrangement events. If choice of a single receptor gene involved gene conversion into a single active locus, cells expressing the P2-IRES-GFP allele might contain a second copy of this allele at the active locus. Southern blotting to distinguish the modified P2-IRES-GFP allele from the unmodified endogenous P2 allele in DNA from control and cloned mice revealed only two bands of equal intensity, suggesting that gene conversion into an active locus does not accompany olfactory receptor gene choice (Supplementary Fig. S2). In addition, Southern blot analyses using multiple DNA probes to examine the organization of about 60 kilobases (kb) of DNA 5' and 3' of the P2 coding sequence failed to detect any differences between donor, control and cloned animals (Supplementary Fig. S3). Sequencing of 10 kb 3' of the P2 coding sequence showed that the P2-IRES-GFP allele was identical in donor, control and cloned mice (data not shown). These results suggest that irreversible changes in DNA do not accompany choice of the P2 odorant receptor gene.

Totipotency of neuronal nuclei

The two-step cloning procedure used to produce mice from neuronal nuclei generates mice in which the neuronally derived ES cells give rise to all embryonic tissues, whereas cells from the tetraploid host blastocyst contribute to the embryonic trophoblast³⁰. Thus, neither this work nor the cloning of lymphocytes via an ES cell intermediate²⁸ demonstrated the totipotency of a nucleus from a terminally differentiated cell³⁴. To demonstrate totipotency of mature OSN nuclei, we transplanted nuclei from P2SN1 and P2SN2 ES cells into enucleated oocytes²⁹. The resulting embryos were cultured for 24 h and transferred to pseudopregnant recipients (Table 1). Upon caesarean section of the recipients, we recovered full-term pups from both the P2SN1 and P2SN2 cell lines. These pups had enlarged placentas (P2SN1, 0.35 g; P2SN2, 0.40 g) but displayed no overt anatomical or behavioural abnormalities, were fertile and survived to adulthood, consistent with previous cloning experiments²⁹. These observations demonstrate that nuclei of terminally differentiated olfactory neurons can be reprogrammed to totipotency, directing development of both embryonic and extra-embryonic lineages.

Discussion

We have asked whether the nucleus of a post-mitotic olfactory sensory neuron can re-enter the cell cycle and undergo reprogramming to direct development of a mouse. ES cell lines were generated from OSN nuclei at frequencies similar to those obtained with differentiated lymphoid cells that can be induced to proliferate under physiological conditions⁷. Thus the mechanisms that lead to the cell-cycle exit and irreversible mitotic arrest that accompany neural differentiation do not result from irreversible epigenetic or genetic events that would interfere with nuclear totipotency.

The differentiation of neurons requires that neural progenitors exit the cell cycle before a restriction point late in G1. This decision is governed by a complex balance of cell-cycle regulators and proneural genes that drive cells into G0 and prevent them from progressing beyond the restriction point³⁵. Although most cells that enter G0 can re-enter the cycle on appropriate stimulation, neurons normally undergo an irreversible mitotic arrest. Whatever mechanisms keep neurons in a post-mitotic state, our experiments demonstrate that they can be overcome in the environment of the egg.

The nervous system contains a diverse array of neural cell types and this diversity is reflected by distinct patterns of gene expression in different neurons. The regulation of gene expression by DNA

rearrangements is rare but this mechanism has nonetheless been suggested to explain the diversity inherent in complex nervous systems³⁶. DNA recombination events provide *Saccharomyces cerevisiae*, trypanosomes and lymphocytes with a mechanism to stochastically express one member of a set of genes that modulate cellular interactions with the environment. One attractive feature shared by gene rearrangements in trypanosomes and lymphocytes is that gene choice by recombination is a random event. Cells that undergo correct or successful rearrangements are then afforded a selective survival advantage. The stochastic rearrangement of one gene from a gene family and subsequent selection could also provide a mechanism to generate the vast diversity of neuronal cell types. In this manner, neurons with subtly different genotypes would exhibit the array of neuronal phenotypes required for a functioning nervous system.

The olfactory sensory epithelium provides a clear example of neuronal diversity, and it has been suggested that this diversity is generated by stochastic DNA rearrangement events^{19,20}. However, efforts to examine the DNA of OSNs expressing a given receptor have been seriously hampered by the inability to obtain homogeneous populations of neurons or clonal cell lines in which each cell expresses the same receptor. We have addressed this problem by generating cloned ES cell lines and mice derived from the nuclei of olfactory sensory neurons expressing the P2 receptor. Analyses of the sequence and organization of the DNA surrounding the expressed P2 allele from cloned ES cells and mice revealed no evidence for either gene conversion or local transpositions at the P2 locus. These results concur with fluorescence *in situ* hybridization studies showing that gene conversion into an active locus is an unlikely mechanism for odorant receptor expression³⁷. In addition, the pattern of receptor gene expression in the sensory epithelium of cloned mice was wild type; multiple odorant receptor genes are expressed without preference for the P2 allele expressed in the donor nucleus.

These data demonstrate that the mechanism responsible for the choice of a single odorant receptor gene does not involve irreversible changes in DNA. More dynamic, reversible recombination events might accompany odorant receptor gene choice, but this is unlikely given the current data. Our results, in combination with experiments showing that the expression of various odorant receptor transgenes is influenced by proximal sequence elements, are consistent with epigenetic models of odorant receptor choice^{37–42}. In a broader context, the generation of fertile cloned mice that are anatomically and behaviourally indistinguishable from wild-type mice indicates that olfactory sensory neurons do not undergo other irreversible DNA rearrangements that would interfere with either the development or function of the nervous system during adult life. □

Methods

Preparation of donor neurons

Olfactory epithelia were dissected in L15 medium (Gibco) at 4 °C, then chopped into small pieces and incubated with type IV collagenase at 1 µg ml⁻¹ at 37 °C for 15 min with occasional vigorous shaking. The collagenase digestion was stopped by addition of DMEM plus 10% FBS. Cells were spun down and resuspended in trituration medium (PBS plus 30% glucose plus 10% FBS plus penicillin/streptomycin) and triturated with several widths of pipette tips to produce single cell suspensions. Cells were pelleted and resuspended in L15/10% FBS before picking.

ES cell lines and tetraploid embryo complementation

Production of ES cell lines from nuclear transfer embryos was exactly as described²⁸, and generation of mice by tetraploid embryo complementation was exactly as described²⁹.

Generation of cloned embryos and mice

Production of cloned embryos by nuclear transfer was essentially as described^{25,29} except that only GFP⁺ cells from the two donor populations, as identified by epifluorescence, were picked and used for nuclear transfer.

In situ hybridization and immunohistochemistry

Animals were killed by approved methods and tissues of interest were either fresh frozen in

OCT or fixed for 2–12 h in 4% PFA/1 × PBS, then washed and equilibrated in 30% sucrose before freezing. Frozen sections (from 20–30 μM) were placed on slides and standard immunohistochemistry and digoxigenin-labelled probe *in situ* hybridization protocols were used³². We used the following antibodies: rabbit anti-GFP antibody (Molecular Probes), rabbit anti-Cre recombinase antibody (Novagen), goat anti-Ki67 antibody (Santa Cruz Biotechnology), rat anti-BrdU antibody (abcam) and rabbit anti-MASH-1 antibody (a gift of J. Johnson).

BrdU labelling and visualization

Mice were injected intraperitoneally with a 5 mg ml⁻¹ solution of BrdU in PBS every 2 h for 12 h. Each mouse received BrdU at 100 μg g⁻¹ body weight for each injection. Two hours after the last injection mice were killed and tissues were fixed for 2 h in 4% PFA and PBS, washed and sucrose protected, and frozen. To permit simultaneous visualization of BrdU and GFP, sections were first stained with anti-GFP antibody then fixed in 4% PFA/PBS for 15 min, then washed, treated with 4 M HCl/0.1% Triton X-100 for 10 min, washed and stained with rat anti-BrdU followed by the appropriate secondary antibodies.

Gene targeting and generation of donor mouse lines

The OMP-IRES-Cre mouse line was generated as described previously except that the Cre recombinase gene sequence was inserted in place of the tTA sequence³². An IRES directing the translation of Cre recombinase was introduced into the 3' untranslated region of the OMP locus by homologous recombination in ES cells. Transgenic mice were generated and crossed with a strain bearing the Z/EG reporter transgene. The Rosa26-LoxP-Stop-LoxP-GFP line was generated as described previously except that the GFP sequence replaced the CFP and YFP sequences in the Rosa26 cassette³³.

Southern blotting, PCR and RT-PCR

Southern blots and PCR screening of ES cells and tails used standard methods⁴³. RT-PCR was performed on RNA isolated from mouse olfactory turbinates that had been frozen in Trizol (Invitrogen), and then treated as per the manufacturer's protocol. RNA was reverse transcribed using the Superscript II kit (Invitrogen) and complementary DNA was subjected to PCR using standard conditions.

Received 13 November 2003; accepted 14 January 2004; doi:10.1038/nature02375.

Published online 15 February 2004.

- Rideout, W. M., Eggan, K. & Jaenisch, R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* **293**, 1093–1098 (2001).
- Hochedlinger, K. & Jaenisch, R. Nuclear transplantation: lessons from frogs and mice. *Curr. Opin. Biol. Sci.* **14**, 741–748 (2002).
- Gurdon, J. B. & Byrne, J. A. The first half-century of nuclear transplantation. *Proc. Natl Acad. Sci. USA* **100**, 8048–8052 (2003).
- Osada, T., Kusakabe, H., Akutsu, H., Yagi, T. & Yanagimachi, R. Adult murine neurons: their chromatin and chromosome changes and failure to support embryonic development as revealed by nuclear transfer. *Cytogenet. Genome Res.* **97**, 7–12 (2002).
- Yamazaki, Y. *et al.* Assessment of the developmental totipotency of neural cells in the cerebral cortex of mouse embryo by nuclear transfer. *Proc. Natl Acad. Sci. USA* **98**, 14022–14026 (2001).
- Chien, Y. H., Gascoigne, N. R., Kavalier, J., Lee, N. E. & Davis, M. M. Somatic recombination in a murine T-cell receptor gene. *Nature* **309**, 322–326 (1984).
- Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl Acad. Sci. USA* **73**, 3628–3632 (1976).
- Brack, C., Hiram, M., Lenhard-Schuller, R. & Tonegawa, S. A complete immunoglobulin gene is created by somatic recombination. *Cell* **15**, 1–14 (1978).
- Gao, Y. *et al.* A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* **95**, 891–902 (1998).
- Frank, K. M. *et al.* Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396**, 173–177 (1998).
- Rehen, S. K. *et al.* Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proc. Natl Acad. Sci. USA* **98**, 13361–13366 (2001).
- Kaushal, D. *et al.* Alteration of gene expression by chromosome loss in the postnatal mouse brain. *J. Neurosci.* **23**, 5599–5606 (2003).
- Chun, J. & Schatz, D. G. Rearranging views on neurogenesis: neuronal death in the absence of DNA end-joining proteins. *Neuron* **22**, 7–10 (1999).
- Chun, J. Selected comparison of immune and nervous system development. *Adv. Immunol.* **77**, 297–322 (2001).
- Yagi, T. Diversity of the cadherin-related neuronal receptor/protocadherin family and possible DNA rearrangement in the brain. *Genes Cells* **8**, 1–8 (2003).
- Zhang, X. & Firestein, S. The olfactory receptor gene superfamily of the mouse. *Nature Neurosci.* **5**, 124–133 (2002).
- Ressler, K. J., Sullivan, S. L. & Buck, L. B. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* **73**, 597–609 (1993).
- Vassar, R., Ngai, J. & Axel, R. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell* **74**, 309–318 (1993).
- Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).
- Kratz, E., Dugas, J. C. & Ngai, J. Odorant receptor gene regulation: implications from genomic organization. *Trends Genet.* **18**, 29–34 (2002).
- Goldstein, B. J. & Schwob, J. E. Analysis of the globose basal cell compartment in rat olfactory epithelium using GBC-1, a new monoclonal antibody against globose basal cells. *J. Neurosci.* **16**, 4005–4016 (1996).
- Holbrook, E. H., Szumowski, K. E. & Schwob, J. E. An immunochemical, ultrastructural, and developmental characterization of the horizontal basal cells of rat olfactory epithelium. *J. Comp. Neurol.* **363**, 129–146 (1995).
- Ohta, Y. & Ichimura, K. Proliferation markers, proliferating cell nuclear antigen, Ki67, 5-bromo-2'-deoxyuridine, and cyclin D1 in mouse olfactory epithelium. *Ann. Otol. Rhinol. Laryngol.* **109**, 1046–1048 (2000).
- Gerdes, J. *et al.* Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody Ki-67. *J. Immunol.* **133**, 1710–1715 (1984).
- Wakayama, T., Perry, A. C., Zuccotti, M., Johnson, K. R. & Yanagimachi, R. Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature* **394**, 369–374 (1998).
- Wakayama, T., Rodriguez, I., Perry, A. C., Yanagimachi, R. & Mombaerts, P. Mice cloned from embryonic stem cells. *Proc. Natl Acad. Sci. USA* **96**, 14984–14989 (1999).
- Rideout, W. M. *et al.* Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nature Genet.* **24**, 109–110 (2000).
- Hochedlinger, K. & Jaenisch, R. Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature* **415**, 1035–1038 (2002).
- Eggan, K. *et al.* Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation. *Proc. Natl Acad. Sci. USA* **98**, 6209–6214 (2001).
- Nagy, A. *et al.* Embryonic stem cells alone are able to support fetal development in the mouse. *Development* **110**, 815–821 (1990).
- Wang, F., Nemes, A., Mendelsohn, M. & Axel, R. Odorant receptors govern the formation of a precise topographic map. *Cell* **93**, 47–60 (1998).
- Gogos, J. A., Osborne, J., Nemes, A., Mendelsohn, M. & Axel, R. Genetic ablation and restoration of the olfactory topographic map. *Cell* **103**, 609–620 (2000).
- Srinivas, S. *et al.* Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Dev. Biol.* **1**, 4 (2001).
- Rossant, J. A monoclonal mouse? *Nature* **415**, 967–969 (2002).
- Ohnuma, S. & Harris, W. A. Neurogenesis and the cell cycle. *Neuron* **40**, 199–208 (2003).
- Edelman, G. M. *Neural Darwinism* (Basic Books, New York, NY, 1987).
- Ishii, T. *et al.* Monoallelic expression of the odorant receptor gene and axonal projection of olfactory sensory neurones. *Genes Cells* **6**, 71–78 (2001).
- Serizawa, S. *et al.* Negative feedback regulation ensures the one receptor-one olfactory neuron rule in mouse. *Science* **302**, 2088–2094 (2003).
- Serizawa, S. *et al.* Mutually exclusive expression of odorant receptor transgenes. *Nature Neurosci.* **3**, 687–693 (2000).
- Ebrahimi, F. A., Edmondson, J., Rothstein, R. & Chess, A. YAC transgene-mediated olfactory receptor gene choice. *Dev. Dyn.* **217**, 225–231 (2000).
- Qasba, P. & Reed, R. R. Tissue and zonal-specific expression of an olfactory receptor transgene. *J. Neurosci.* **18**, 227–236 (1998).
- Vassalli, A., Rothman, A., Feinstein, P., Zapotocky, M. & Mombaerts, P. Minigenes impart odorant receptor-specific axon guidance in the olfactory bulb. *Neuron* **35**, 681–696 (2002).
- Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: a Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Moring, A. Nemes, M. Mendelsohn, J. Loring, J. Dausman, A. Meissner and K. Hochedlinger for assistance during the course of these experiments; T. Cutforth, J. de Nooij, T. Jessell and J. Johnson for sharing reagents necessary for our experiments; and members of the Jaenisch, Axel and Chess laboratories for discussion and assistance during the course of these experiments, especially F. A. Ebrahimi, M. Rios, W. M. R. Rideout, L. Jackson-Grusby and B. Shykind. This research was sponsored by NIH grants to R.J., R.A., A.C. and M.T. K.B. is an Associate and R.A. is an Investigator of the Howard Hughes Medical Institute. K.E. is a Junior Fellow in the Harvard Society of Fellows.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.J. (jaenisch@wi.mit.edu) or A.C. (achess@wi.mit.edu).

Cavity cooling of a single atom

P. Maunz, T. Puppe, I. Schuster, N. Syassen, P. W. H. Pinkse & G. Rempe

Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Str. 1,
D-85748 Garching, Germany

All conventional methods to laser-cool atoms rely on repeated cycles of optical pumping and spontaneous emission of a photon by the atom. Spontaneous emission in a random direction provides the dissipative mechanism required to remove entropy from the atom. However, alternative cooling methods have been proposed^{1,2} for a single atom strongly coupled to a high-finesse cavity; the role of spontaneous emission is replaced by the escape of a photon from the cavity. Application of such cooling schemes would improve the performance of atom-cavity systems for quantum information processing^{3,4}. Furthermore, as cavity cooling does not rely on spontaneous emission, it can be applied to systems that cannot be laser-cooled by conventional methods; these include molecules² (which do not have a closed transition) and collective excitations of Bose condensates⁵, which are destroyed by randomly directed recoil kicks. Here we demonstrate cavity cooling of single rubidium atoms stored in an intracavity dipole trap. The cooling mechanism results in extended storage times and improved localization of atoms. We estimate that the observed cooling rate is at least five times larger than that produced by free-space cooling methods, for comparable excitation of the atom.

The basic idea behind cavity cooling can be understood from a simple classical picture based on the notion of a refractive index. Consider a standing-wave optical cavity resonantly excited by a weak probe laser blue detuned from the atomic resonance. For strong atom-cavity coupling, even one atom can significantly influence the optical path length between the cavity mirrors. Consequently, the intracavity intensity is strongly affected by the atom^{6–8}. For example, at a node of the standing wave the atom is not coupled to the cavity, and thus the intracavity intensity is large. An atom at an antinode, in contrast, shifts the cavity to a higher frequency because the atom's refractive index is smaller than unity above its resonance. This tunes the cavity out of resonance from the probe laser and leads to a small intracavity intensity. However, in a high-finesse cavity the intensity cannot drop instantaneously when the atom moves away from a node. Instead, the blue-shift of the cavity frequency leads to an increase of the energy stored in the field.

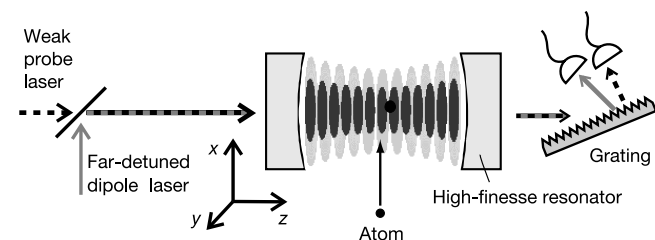


Figure 1 Experimental set-up. The high-finesse cavity ($F = 4.4 \times 10^5$) is excited by a weak near-resonant probe field and a strong far-red detuned dipole field. ^{85}Rb atoms are injected from below. Behind the cavity, the two light fields are separated by a grating. The probe light is further passed through a narrow-band interference filter before being directed onto a single-photon counting module. For this set-up, a quantum efficiency of 32% is achieved for the detection of probe light transmitted through the cavity while the dipole light is attenuated by more than 70 dB. The dipole light is also used to stabilize the cavity length with a radio frequency sideband technique. It is generated by a grating- and current-stabilized diode laser with a linewidth of 20 kHz r.m.s..

The photons finally escaping from the cavity are therefore blue-shifted from the photons of the probe laser. This occurs at the expense of the atom's kinetic energy. The reverse effect, namely the acceleration of an atom approaching an antinode, is much smaller, as here the cavity is initially out of resonance with the probe laser and consequently the intracavity intensity is small.

Note that the cooling process does not require atomic excitation. Indeed, the atomic excitation is low at all times, as the atom is not coupled to the light at a node while the intracavity intensity is very low for an atom near an antinode. It follows that the lowest attainable temperature is not limited by the atomic linewidth as for free-space Doppler cooling but by the linewidth of the cavity, which can be much smaller. Therefore temperatures below the Doppler limit can be reached⁹. An upper limit on the velocity of the atom to be cooled is given by the requirement that the atom must not move farther than about one-quarter of a wavelength during the lifetime of a photon in the cavity. In our experiment, this corresponds to a velocity of about 3 m s^{-1} . We emphasize that cavity cooling is applied to a single two-level atom and differs from the mechanical effects observed for an atomic ensemble^{10–12}. A description of cavity cooling in terms of dressed-states pictures of the strongly coupled atom-cavity system can be found elsewhere⁹.

A treatment of cavity cooling combined with trapping by means of an auxiliary far-red detuned dipole laser is quantitatively different. It can be achieved by including the dynamic Stark shift of the atomic ground and excited states in the hamiltonian. The dynamic Stark shift renders the atomic resonance frequency position dependent, making it larger for an atom at an antinode. This effect even enhances the cooling force by effectively increasing the refractive index variations for a moving atom. The combined system can be investigated numerically¹³. Moreover, in the limit of low atomic excitation analytic expressions for all relevant forces including the cooling force can be derived, thereby extending previous calculations^{1,9}. The obtained expressions are lengthy but valuable for parameter optimization and straightforward trajectory calculation.

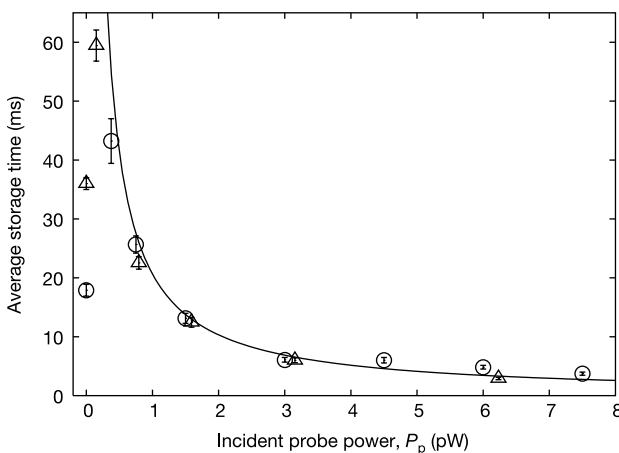


Figure 2 Storage time. The storage time of a trapped atom as a function of probe power. Two sets of data taken before (circles) and after (triangles) improving the laser stabilization are displayed. The corresponding storage times in the dark trap are 18 ms and 36 ms, respectively. For an incident probe power exceeding about 0.5 pW, the storage time is limited by radial escape due to spontaneous emission. In this range the measured storage times (in ms) can be approximated by $20/P_p$ (where P_p is in units of pW; solid line). The storage time of 18 ms obtained for the dark trap is increased by a factor of more than two by applying 0.37 pW probe light. After the stabilization of the dipole laser was improved, $P_p = 0.11 \text{ pW}$ increased the storage time from 36 ms to 60 ms. The mentioned probe powers correspond to an average intracavity photon number of 0.005 and 0.0015 for 0.37 pW and 0.11 pW, respectively.

In our experiment (Fig. 1), the cavity has a finesse $\mathcal{F} = 4.4 \times 10^5$ and a length $l = 120 \mu\text{m}$. The cavity field has a decay rate $\kappa/2\pi = 1.4 \text{ MHz}$. The wavelength difference between neighbouring longitudinal TEM₀₀ modes is about 2.5 nm. Single laser-cooled ^{85}Rb atoms are injected into the cavity⁷ with a velocity smaller than 10 cm s^{-1} . The single-photon coupling constant is $g/2\pi = 16 \text{ MHz}$ for the $5^2S_{1/2}F=3 \leftrightarrow 5^2P_{3/2}F=4$ transition with dipole decay rate $\gamma/2\pi = 3 \text{ MHz}$. A weak, near-resonant probe laser at 780.2 nm is used to observe and cool the atom. A strong, far-detuned dipole laser at 785.3 nm serves to trap the atom. The detunings of the probe laser with respect to the cavity, $\Delta_c = 0$, and the atom, $\Delta_a/2\pi = 35 \text{ MHz}$, are chosen as a compromise between ideal detunings for detection and good cooling conditions while maintaining a very low excitation of the atom at any moment of the experiment. In fact, the presence of an atom at an antinode reduces the transmission of the probe light by a factor of 100, allowing its detection and manipulation with a high signal-to-noise ratio and a high bandwidth^{14–17}.

Experiments are performed only with atoms located in the central region of the cavity, where the nodes and antinodes of the probe and dipole fields coincide. This is accomplished by turning on the dipole laser before injecting the atom into the cavity. In this case, a 400 μK deep dipole potential guides the arriving atom into the high-intensity region. Only if this region is also a region of strong coupling does the presence of an atom manifest itself as a pronounced dip in the cavity transmission as monitored with the probe light. Atoms entering the cavity at an axial position where the two standing waves are out-of-phase are confined to nodes of the probe field and, hence, are invisible to the probe laser. If the transmission drops below 9%, the intensity of the dipole light is increased to generate a trap depth of about 1.5 mK. This compensates for the radial kinetic energy of the atom and enables the atom to be caught in an otherwise conservative potential. More than 95% of the detected atoms are captured in the dipole trap.

The average storage time of the atom in the dark intracavity trap is measured by turning off the probe light for an adjustable time interval, Δt , after the atom is captured. As a function of the dark time, Δt , the fraction of atoms still trapped drops exponentially with a decay constant of 18 ms, defining the storage time of the atoms in

the trap. The theoretical limits imposed by light scattering (85 s) and genuine cavity quantum electrodynamic dipole fluctuations (200 ms)¹ cannot explain this rather short time^{18,19}. Instead, the observed loss of the atom is attributed to parametric heating due to fluctuations of the intracavity intensity mainly caused by frequency fluctuations of the dipole laser. This technical noise depends critically on the tuning of the laser stabilization. Indeed, the storage time in the dark dipole trap could be increased from 18 ms to 31 ms by improving the frequency stability of the dipole laser. As this stability and other sources of noise are hard to control, a concurrent measurement of the storage time of the dark dipole trap serves as a reference in each of the following experiments. Note that the axial trap frequency is about 100 times higher than the radial trap frequency. As parametric heating is proportional to the square of the trap frequency²⁰ the atom is mainly heated in the axial direction. Because axial and radial motion are only weakly coupled, the heated atom usually escapes the antinode of the standing wave dipole trap along the axis, thereby hitting one of the mirrors. This conjecture is supported by numerical simulations of the experiment, as further discussed below.

To demonstrate that cavity cooling can be used to compensate for the axial heating of the dipole trap, the probe beam is not switched off completely after capturing an atom. Figure 2 shows the storage time as a function of the incident probe power, P_p . For high probe power, the storage time is reduced as compared to the dark trap. However, the storage time increases with decreasing power. This effect is attributed to the reduction of spontaneous emission, which heats the atom in all directions and which cannot be compensated radially because cavity cooling acts mainly axially. As the atomic transition is still far from being saturated, the radial heating is proportional to the probe power even for the highest considered level of $P_p = 7.5 \text{ pW}$. Hence, as long as the probe power is large enough to compensate for axial heating, the storage time τ is limited by radial loss and $\tau \propto P_p^{-1}$ (solid line in Fig. 2). Consequently, the atom is expected to leave the cavity axially for near-zero probe power, whereas for higher power radial losses should dominate. This is confirmed by a Monte Carlo simulation of a point-like atom

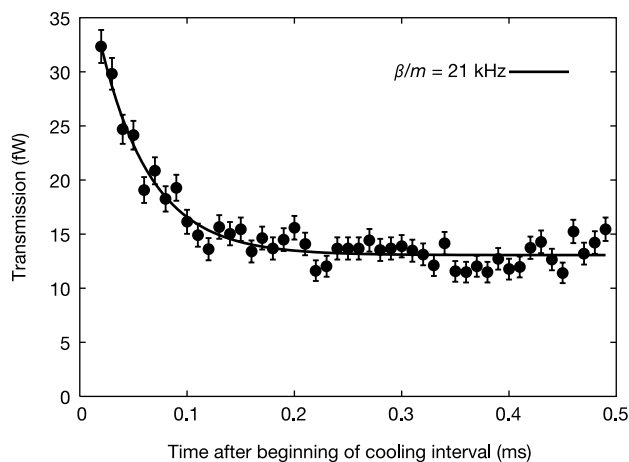


Figure 3 Cooling-induced localization. Average transmission during cooling intervals of length 0.5 ms after heating the atom for 0.1 ms. An incident power of $P_p = 2.25 \text{ pW}$ is chosen for a good signal to noise ratio. Without an atom, the cavity transmission on resonance is 300 fW. The atoms are cooled during the first 0.1 ms. This leads to a stronger coupling to the cavity mode and, hence, to a smaller transmission. A cooling rate of $\beta/m = 21 \text{ kHz}$ is estimated from an exponential fit. Radial heating occurs on a much longer timescale and is not visible here.

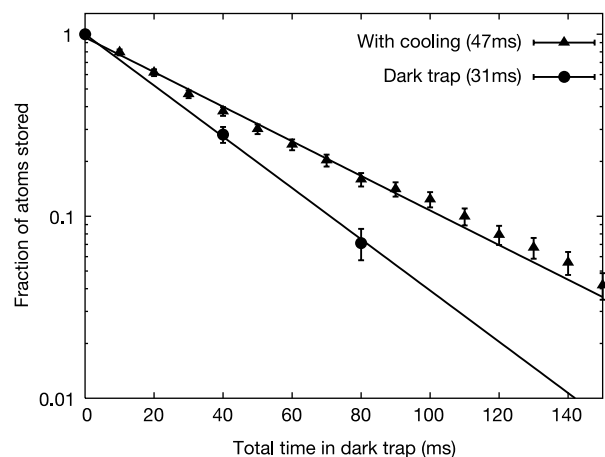


Figure 4 Cavity cooling. The fraction of atoms stored in the trap as a function of time after the dipole trap is switched on. The dark dipole trap has an average storage time of 31 ms (circles), as obtained from an exponential fit. If 100- μs -long cooling intervals are applied every 2 ms, the average storage time without counting the cooling intervals is extended to 47 ms (triangles). The storage time in the dark trap varies slightly from day to day and is therefore measured concurrently. As in several detailed measurements (not shown) the fraction of atoms found in the dark trap as a function of time was found to be very well described by an exponential decay; three data points (including the one at zero trapping time) are sufficient to obtain the storage time in the dark trap.

moving in the trap under the influence of the forces and momentum diffusion calculated analytically, while parametric heating from the dipole trap is implemented by a randomly changing potential depth. The storage times evaluated from the simulation agree well with the experiment. Moreover, it can be concluded that for probe powers below 0.1 pW more than 90% of the atoms leave the cavity by hitting a mirror, whereas for higher probe powers 90% of the atoms leave radially.

We emphasize that, in contrast to Doppler cooling, cavity cooling extends the storage time for a probe field which is blue-detuned from the atom. If the detuning is changed from blue to red by adjusting the atom–cavity detuning while keeping the dipole power constant and the probe laser resonant with the cavity, the average storage time decreases and drops below the storage time of the dark trap. This clearly demonstrates that the extension of the storage time cannot be attributed to Doppler cooling.

We now demonstrate cooling by directly observing the reduction of the kinetic energy of the atom. The experiment employs alternating heating and cooling intervals at a constant probe power of $P_p = 2.25$ pW. Axial heating is achieved by deliberately tuning the probe laser for a time interval of 100 μ s to a frequency 9 MHz above the cavity resonance ($\Delta_c/2\pi = 9$ MHz), where strong heating of the atom is expected from the theoretical analysis. In the following 500 μ s-long cooling interval, the probe laser is switched back to the cavity resonance ($\Delta_c = 0$). Figure 3 shows the cavity transmission averaged over many cooling intervals, with an atom present in the cavity. The transmission drops by a factor of more than two during the first 100 μ s. This drop is a clear signature of increasing atom–cavity coupling, and so indicates better localization of the atom at the antinode.

The exponential change of the transmission observed for short times allows us to obtain an estimate of the mean cooling rate, $\beta/m = 21$ kHz, where β is the friction coefficient and m the atomic mass. This result is in good agreement with the Monte Carlo simulations, which show an exponential relaxation with a timescale of about 50 μ s for the increase of the atomic localization as well as for the decrease of the transmitted power. To compare this rate of cavity cooling with free-space cooling rates of a two-level atom for a given rate of spontaneous emission events, knowledge about the atomic excitation is required. Here, an upper limit can be obtained by attributing the storage time of 9 ms (measured for a probe power of 2.25 pW; Fig. 2) solely to radial heating from spontaneous emission. To leave the trap, an atom must have gained about 1 mK of kinetic energy. This limits the atomic excitation to below 2.5%. At this excitation, free-space Sisyphus cooling^{21,22} of a two-level atom in a blue-detuned standing wave would achieve $\beta_s/m = 4$ kHz while Doppler cooling would have $\beta_D/m = 1.5$ kHz, both for optimal detuning. Thus introducing the cavity increases the cooling rate by a factor of at least 5 for constant atomic excitation.

Cooling down an atom in the trap can also be demonstrated without the additional heating. For this purpose, the atom is repeatedly left in the dark dipole trap by switching off the probe light for 2-ms-long time intervals. These intervals are short compared to the average trapping time in the dark trap of 31 ms, obtained after improving the frequency stability of the dipole laser. In each of the dark intervals the atom experiences parametric heating. Between the dark intervals, 100- μ s-long cooling intervals are applied using a probe power $P_p = 1.5$ pW on resonance with the cavity ($\Delta_c = 0$). The transmission of the probe light is also used to determine whether the atom is still present. The average time an atom is stored in the trap under these conditions can be calculated by adding up all the 2-ms-long dark intervals after which the atom is still found in the trap, but omitting the cooling intervals. The result is shown in Fig. 4: although the short cooling intervals have a duty cycle of only 5% they increase the average trapping time by more than 50%. Obviously, heating the atom out of the trap requires

more time in the presence of cooling. Therefore the kinetic energy of the atom was reduced during the cooling interval.

We have used strong coupling of a two-level atom to the cavity field to cool single atoms stored in an intracavity dipole trap. The storage time in the trap has been increased by a factor of two by exploiting the cooling force caused by a near-resonant cavity field with an average photon number of only 0.005. In contrast to free-space laser-cooling techniques, this cooling force acts mainly by exciting the cavity part of the coupled atom–cavity system. Thus strong cooling forces can be achieved while keeping the atomic excitation low. An estimate of the strength of the cooling force has shown to exceed the force expected for free-space Sisyphus cooling and Doppler cooling at comparable atomic excitation by at least a factor of 5 and 14, respectively. Avoiding excitation could serve as a basis for cooling of molecules or collective excitations of a Bose condensate⁵. Another application might be to cool the motion of an atom with a stored quantum bit²³. If the two states forming the qubit had identical coupling to the cavity, the cooling scheme would not disturb the superposition state. This advantage is not shared by any other cooling method. $\square$

Received 9 December 2003; accepted 4 February 2004; doi:10.1038/nature02387.

- Horak, P., Hechenblaikner, G., Gheri, K. M., Stecher, H. & Ritsch, H. Cavity-induced atom cooling in the strong coupling regime. *Phys. Rev. Lett.* **79**, 4974–4977 (1997).
- Vuletić, V. & Chu, S. Laser cooling of atoms, ions, or molecules by coherent scattering. *Phys. Rev. Lett.* **84**, 3787–3790 (2000).
- Kuhn, A., Hennrich, M. & Rempe, G. Deterministic single-photon source for distributed quantum networking. *Phys. Rev. Lett.* **89**, 067901 (2002).
- Monroe, C. Quantum information processing with atoms and photons. *Nature* **416**, 238–246 (2002).
- Horak, P. & Ritsch, H. Dissipative dynamics of Bose condensates in optical cavities. *Phys. Rev. A* **63**, 023603 (2001).
- Mabuchi, H., Turchette, Q. A., Chapman, M. S. & Kimble, H. J. Real-time detection of individual atoms falling through a high-finesse optical cavity. *Opt. Lett.* **21**, 1393–1395 (1996).
- Münstermann, P., Fischer, T., Pinkse, P. W. H. & Rempe, G. Single slow atoms from an atomic fountain observed in a high-finesse optical cavity. *Opt. Commun.* **159**, 63–67 (1999).
- Sauer, J. A., Fortier, K. M., Chang, M. S., Hamley, C. D. & Chapman, M. S. Cavity QED with optically transported atoms. Preprint at (<http://arXiv.org/quant-ph/0309052>) (2003).
- Hechenblaikner, G., Gangl, M., Horak, P. & Ritsch, H. Cooling an atom in a weakly driven high-Q cavity. *Phys. Rev. A* **58**, 3030–3042 (1998).
- Chan, H. W., Black, A. T. & Vuletić, V. Observation of collective-emission-induced cooling of atoms in an optical cavity. *Phys. Rev. Lett.* **90**, 063003 (2003).
- Nagornyi, B., Elsässer, Th. & Hemmerich, A. Collective atomic motion in an optical lattice formed inside a high finesse cavity. *Phys. Rev. Lett.* **91**, 153003 (2003).
- Kruse, D., von Cube, C., Zimmermann, C. & Courtillot, Ph. W. Observation of lasing mediated by collective atomic recoil. Preprint at (<http://arXiv.org/quant-ph/0305033>) (2003).
- van Enk, S. J., McKeever, J., Kimble, H. J. & Ye, J. Cooling of a single atom in an optical trap inside a resonator. *Phys. Rev. A* **64**, 013407 (2001).
- Hood, C. J., Lynn, T. W., Doherty, A. C., Parkins, A. S. & Kimble, H. J. The atom-cavity microscope: single atoms bound in orbit by single photons. *Science* **287**, 1447–1453 (2000).
- Pinkse, P. W. H., Fischer, T., Maunz, P. & Rempe, G. Trapping an atom with single photons. *Nature* **404**, 365–368 (2000).
- Münstermann, P., Fischer, T., Maunz, P., Pinkse, P. W. H. & Rempe, G. Dynamics of single-atom motion observed in a high-finesse cavity. *Phys. Rev. Lett.* **82**, 3791–3794 (1999).
- Fischer, T., Maunz, P., Pinkse, P. W. H., Puppe, T. & Rempe, G. Feedback on the motion of a single atom in an optical cavity. *Phys. Rev. Lett.* **88**, 163002 (2002).
- Ye, Y., Vernooy, D. W. & Kimble, H. J. Trapping of single atoms in cavity QED. *Phys. Rev. Lett.* **83**, 4987–4990 (1999).
- McKeever, J. et al. State-insensitive cooling and trapping of single atoms in an optical cavity. *Phys. Rev. Lett.* **90**, 133602 (2003).
- Savard, T. A., O'Hara, K. M. & Thomas, J. E. Laser-noise-induced heating in far-off resonance optical traps. *Phys. Rev. A* **56**, R1095–R1098 (1997).
- Aspect, A., Dalibard, J., Heidmann, A., Salomon, C. & Cohen-Tannoudji, C. Cooling atoms with stimulated emission. *Phys. Rev. Lett.* **57**, 1688–1691 (1986).
- Dalibard, J. & Cohen-Tannoudji, C. Dressed-atom approach to atomic motion in laser light: the dipole force revisited. *J. Opt. Soc. Am. B* **2**, 1707–1720 (1985).
- Griessner, A., Jaksch, D. & Zoller, P. Cavity assisted nondestructive laser cooling of atomic qubits. Preprint (<http://arXiv.org/quant-ph/0311054>) (2003).

Acknowledgements This work was partially funded by the German Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G. R. (gerhard.rempe@mpq.mpg.de).

An explanation for a universality of transition temperatures in families of copper oxide superconductors

Sudip Chakravarty¹, Hae-Young Kee² & Klaus Völker²

¹Department of Physics and Astronomy, University of California Los Angeles, Los Angeles, California 90095-1547, USA

²Department of Physics, University of Toronto, Ontario M5S 1A7, Canada

A remarkable mystery of the copper oxide high-transition-temperature (T_c) superconductors is the dependence of T_c on the number of CuO_2 layers, n , in the unit cell of a crystal. In a given family of these superconductors, T_c rises with the number of layers, reaching a peak at $n = 3$, and then declines¹: the result is a bell-shaped curve. Despite the ubiquity of this phenomenon, it is still poorly understood and attention has instead been mainly focused on the properties of a single CuO_2 plane. Here we show that the quantum tunnelling of Cooper pairs between the layers² simply and naturally explains the experimental results, when combined with the recently quantified charge imbalance of the layers³ and the latest notion of a competing order^{4–9} nucleated by this charge imbalance that suppresses superconductivity. We calculate the bell-shaped curve and show that, if materials can be engineered so as to minimize the charge imbalance as n increases, T_c can be raised further.

The phase diagrams of the high- T_c superconductors show two striking universalities. The first is the dependence of T_c on the number of CuO_2 layers, n , within a homologous series, as shown in Fig. 1. The second is the superconducting dome: as a function of doping, x (charge carriers added to the conducting CuO_2 planes), T_c is a bell-shaped curve rising at about $x \approx 0.05$ and dropping to zero at $x \approx 0.3$; see Fig. 2 for a similar behaviour of the superconducting order parameter.

It is also evident that these superconductors are rife with competing orders, of which the most recent experimental demonstration is the measurement of the Hall number in the 60 T magnetic field¹⁰. A specific suggestion has been that the demise of superconductivity, and hence the universal origin of the superconducting dome as a function of x , is due to a new competing order called the d -density wave (DDW)⁸, where a particle and a hole are bound in the angular momentum channel of $l = 2$ (d -wave). The second universal feature was also recognized previously, and interlayer tunnelling theory was developed to explain it². Unfortunately, this theory leads to a T_c that saturates as a function of n (ref. 11), which is not the observed bell-shaped curve shown in Fig. 1. In addition, a particularly strong conjecture¹² that the entire pairing mechanism was due to interlayer tunnelling was later called into question by experiments¹³.

In contrast to ref. 12, we argue that the tunnelling between the layers must be regarded as a mechanism by which pairing is enhanced^{14,15} and should not be construed as the sole reason for high T_c , especially for single-layer materials in which the effect of tunnelling is negligible. To take advantage of tunnelling between the close pairs of CuO_2 planes within a unit cell, it is necessary that a low-energy electron in the normal state is forbidden to tunnel coherently perpendicular to the planes². In the superconducting state this kinetic energy is recovered, resulting in enhanced pairing. The frustrated kinetic energy in the normal state may either be due to a non-Fermi-liquid nature of this state, in which an electron breaks up into more fundamental constituents², or due to the pseudogap that is present over much of the phase diagram¹⁶. From a simple renormalization group argument, in the pseudogap state, the single-particle tunnelling is irrelevant at low energies

below the gap, and it should simply drop out of all macroscopic considerations¹⁷. In contrast, pair tunnelling, in which a Cooper pair of electrons tunnels together, is a coherent zero-energy process and must have macroscopic consequence. For high- T_c superconductors, the tunnelling matrix element in the momentum space is peaked where the superconducting gap is large², and therefore the existence of nodes in the gap cannot invalidate this argument. The importance of pair tunnelling is also crucial in stripe theories of superconductivity¹⁸.

We shall not dwell further on microscopic considerations, because the important aspect of the phase diagram can be understood using a suitable free-energy functional. Imagine that two-dimensional planes, indexed by j , are stacked to form unit cells consisting of n layers of a three-dimensional superconductor such that the mean-field free-energy functional of the complex order parameters ψ_j in a unit cell is (A is the area of the two-dimensional plane):

$$F_s = A \sum_j [\alpha' |\psi_j|^2 + \lambda' |\psi_j|^4 - \rho_c (\psi_j \psi_{j+1}^* + \text{c.c.})] \quad (1)$$

where α' , λ' , and ρ_c are parameters that may in principle depend on the layer index, and the gradient terms are left out in mean field theory; 'c.c.' stands for complex conjugation. At temperature $T = 0$, F_s is the ground-state energy. On symmetry grounds, there are no distinctions between the following two choices of the coupling between the layers: $-(\psi_j \psi_{j+1}^* + \text{c.c.})$ and $|\psi_j - \psi_{j+1}|^2$. They are simply related by a shift of the quadratic term. Our tacit choice results in an enhancement of pairing, while the choice $|\psi_j - \psi_{j+1}|^2$ does not, because its minimum value is zero¹⁹; the crux, of course, is to enhance the mean-field T_c , as fluctuations can only reduce it. Equation (1) results in a T_c that initially rises with n and then saturates¹¹. Although the rise can be significant, the crucial subsequent drop and the maxima at $n = 3$ are not captured.

We now include the free energy of the competing order parameter that is ultimately responsible for the downturn of T_c with n . For concreteness, consider a model where the competing order is DDW, but any other suitable order²⁰ may also serve the same purpose. For the DDW order parameter ϕ_j that breaks the time reversal symmetry and couples (g is the coupling) to the superconducting order parameter ψ_j , the free energy is:

$$F_c = A \sum_j [\alpha \phi_j^2 + \lambda \phi_j^4 + g |\psi_j|^2 \phi_j^2] \quad (2)$$

The parameters α , λ and g may depend on the layer index. The coupling between ψ and ϕ is a fourth-order invariant, and therefore, to be consistent, the individual fourth-order terms in equations (1)

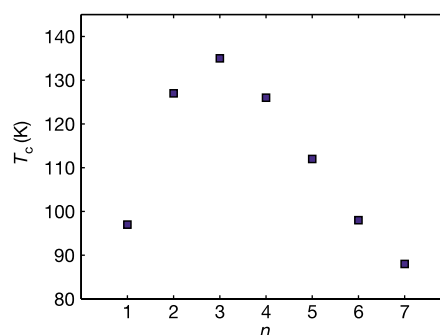


Figure 1 Transition temperature within a homologous series. A homologous series²⁹ is a family of compounds having the same charge-reservoir block, but $n\text{CuO}_2$ -planes in the infinite-layer block, which in turn consists of $(n-1)$ bare cation planes and $n\text{CuO}_2$ -planes. A good example is the family $\text{HgBa}_2\text{Ca}_{(n-1)}\text{CuO}_{(2n+2+\delta)}$ whose T_c as a function of n , optimized with respect to oxygen concentration, is shown. (The figure is adapted from the data in ref. 30.) Similar results have been known for some time; see ref. 1.

and (2) must not be neglected. We have ignored the negligibly small interlayer coupling of the DDW order parameter ϕ_j (ref. 21).

Another important ingredient in our theory is recent nuclear magnetic resonance (NMR) measurements of the Hg-series^{3,22,23}. Using an empirical relation between the spin part of the ⁶³Cu Knight shift and the hole concentration for materials with $n = 1$ and $n = 2$, the doping of the individual planes up to $n = 5$ was deduced. It was found that in a unit cell the outer layers tend to get overdoped, while the inner layers tend to get underdoped, consistent with simple but robust electrostatic Madelung energy considerations³.

The complete Ginzburg–Landau free energy at $T = 0$ of a n -layer system, including the doping imbalance, is given by $F = F_s + F_c$, where:

$$F = A \sum_j \left[\alpha'(x_j) |\psi_j|^2 + \lambda' |\psi_j|^4 - \rho_c (\psi_j \psi_{j+1}^* + \text{c.c.}) + \alpha(x_j) \phi_j^2 + \lambda \phi_j^4 + g |\psi_j|^2 \phi_j^2 \right] \quad (3)$$

and $\alpha'(x_j)$ and $\alpha(x_j)$ are functions of the doping x_j for each layer; all other parameters are assumed to be constants. For a single-layer system, there is only one term in the sum, $x_j \equiv x$, $\rho_c \equiv 0$, and the minimum of the free energy is easily found analytically. The free-energy functional in equation (3) gives the right shape and the magnitude of the superconducting dome for a generic single-layer material, as shown in Fig. 2.

In the general case we have to resort to numerical minimization. We use a downhill simplex method combined with simulated annealing to determine the individual order parameters for each layer in a unit cell, with an open boundary condition. As indicators for the superconducting transition temperature, we have examined the following three criteria: $T_c \propto \psi_{\text{avg}} = \frac{1}{n} \sum_i |\psi_i|$, $T_c \propto \psi_{\text{r.m.s.}} = (\frac{1}{n} \sum_i |\psi_i|^2)^{1/2}$, and $T_c \propto \psi_{\text{max}} = \max |\psi_j|$. The choice between these three indicators is very complex, involving many details related to the proximity effect. In a nutshell, ψ_{avg} is a better indicator when the superconducting coherence length in a direction perpendicular to the planes is relatively large, whereas ψ_{max} is more appropriate in the opposite limit when this length is very short. Fortunately, as we shall see, the robust features of the phase diagram are independent of the choice. We have refrained, however, from using a finite-temperature version of the free-energy functional, because it would have introduced additional adjustable parameters. The parameters in equation (3) are strongly constrained by the physics of the single-layer problem. Figure 3 shows the dependence of T_c on the number of layers. The results capture the dependence found in experiments: interlayer coupling enhances the transition

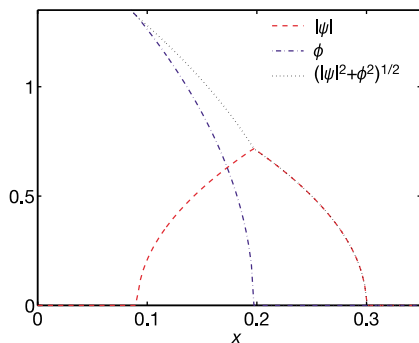


Figure 2 The $T = 0$ phase diagram of a one-layer copper oxide as a function of doping, x . To reproduce the well-known superconducting dome of the order parameter of a single-layer copper oxide, we choose $\alpha(x) = 27(x - 0.22)$, $\alpha'(x) = 10(x - 0.3)$, $\lambda = \lambda' = 1$ and $g = 1.2$. These are the same parameters as those in ref. 8. The order parameter ψ is the superconducting order parameter, and ϕ is the competing order parameter.

temperature up to $n = 3$, but the large doping imbalance, combined with the competing order, reduces the optimal T_c beyond three layers.

According to the theory of competing order⁸, the total single-particle excitation gap, E_g , is not the superconducting gap alone, but a function of both the superconducting gap and the gap due to the hidden order, the DDW gap for instance. Whereas the superconducting gap is controlled by the order parameter, ψ , the competing gap is controlled by ϕ . Thus, E_g can be finite even when ψ is zero. This is different from those theories where the total gap is the local superconducting gap alone, and the decrease in T_c with x is due to phase fluctuations²⁴. Although phase fluctuations do play a role in determining T_c , in our view it is the competition of the two order parameters that plays the dominant role in determining the superconducting dome.

In principle, we can further adjust parameters to fit data better, but it would not be very meaningful for two reasons. First, the actual enhancement of T_c for different homologous series differs in magnitude. Second, we have only calculated the mean-field order parameters at $T = 0$. Fluctuation effects will surely be important for the actual transition temperatures. Fluctuations will depress the T_c for $n = 1$ more than the T_c for $n = 3$. Thus, the overall increase in T_c will be in better agreement with experiments. We note that conventional Josephson coupling between the layers, $|\psi_j - \psi_{j+1}|^2$, can suppress fluctuations and raise the actual transition temperature closer to the mean-field value as the layers are coupled¹⁹. This mechanism was indeed explored²⁵, but the enhancement of T_c was too small to explain the striking data shown in Fig. 1. Moreover, a mechanism to explain the bell-shaped profile of T_c as a function of n was missing in this calculation.

According to our theory, to increase T_c further, it would be necessary to dope the system in such a way that the layers do not develop a charge imbalance and nucleate competing order. Another consequence would be that site-specific NMR relaxation rates should show a pseudogap in the inner layers, but not in the outer

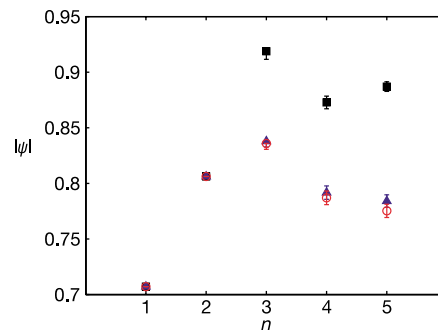


Figure 3 The calculated superconducting order parameters at $T = 0$ of multilayer copper oxides. For $n \geq 3$, the parameters $\alpha(x_j)$ and $\alpha'(x_j)$ in equation (3) vary with the layers because of the charge imbalance between the inner and the outer layers in a unit cell. Once we know the doping of a layer x_j , we use the same parameterizations as in the legend of Fig. 2. Thus, the superconducting dome is entirely determined by the physics of the competing order of a single layer material. In accordance with ref. 3, we assume that an amount of charge ϵ is transferred from the $(n - 2)$ inner layers to the two outer layers, so that the effective doping is $x_1 = [1 - \epsilon/(n - 2)]x$ on the inner layers, and $x_0 = [1 + \epsilon/2]x$ on the outer layers. Here x is the average, or nominal, doping per layer. The charge imbalance ϵ is extracted from ref. 3, where the ratio $x_0/x_1 = R_h$ is given by 1.14 for $n = 3$, 1.49 for $n = 4$, and 1.64 for $n = 5$, with $x \approx 0.2$. Because $\epsilon = [2(n - 2)(R_h - 1)]/(n - 2 + 2R_h)$, we obtain $\epsilon = 0.085$, 0.39 and 0.61 for the 3-, 4- and 5-layer systems, respectively. The interlayer coupling ρ_c is the only remaining free parameter, which we set to be 0.3 to optimize the shape of T_c versus n . Here ψ_{avg} (circles), $\psi_{\text{r.m.s.}}$ (triangles) and ψ_{max} (squares) are shown as a function of the number of layers n , as the indicators for the superconducting transition temperature T_c . The error bars were supplied by Y. Kitaoka (personal communication).

layers, owing to the charge imbalance, which is in agreement with recent experiments^{22,23}. This is because the suppression of the superconducting order parameter in the inner layers will enhance the pseudogap. The single-particle excitation spectra as observed in angle-resolved photoemission spectroscopy (ARPES) of multilayer copper oxides should be sensitive to the doping imbalance of the layers; there is already some indication of this in experiments^{26,27} involving the triple-layer material Bi2223. Although bilayer splitting is observed in optimally doped Bi2212, trilayer and higher-multilayer splittings will be increasingly difficult to observe, because of the induced pseudogap of the inner layers.

We also predict that the maximum superconducting gap measured in ARPES will be a bell-shaped curve as a function of n with a maximum at $n = 3$. It would be worth investigating whether the recently developed Fourier-transform scanning tunnelling spectroscopy (FT-STs)²⁸ could provide layer-specific information. The change in the spectra with increasing n should be detectable, as the tunnelling rate falls off exponentially with the distance. For this, it would be interesting to consider an underdoped sample. As n increases, the spectra, which would be dominated by the outer layer, will change because of its increased doping. □

Received 16 November 2003; accepted 16 January 2004; doi:10.1038/nature02348.

1. Scott, B. A. *et al.* Layer dependence of the superconducting transition temperature of $\text{HgBa}_2\text{Ca}_{n-1}\text{Cu}_n\text{O}_{2n+2+\delta}$. *Physica C* **230**, 239–245 (1994).
2. Chakravarty, S., Sudbo, A., Anderson, P. W. & Strong, S. Interlayer tunneling and gap anisotropy in high-temperature superconductors. *Science* **261**, 337–340 (1993).
3. Kotev, H. *et al.* NMR study of carrier distribution and superconductivity in multilayered high- T_c cuprates. *J. Phys. Chem. Solids* **62**, 171–175 (2001).
4. Perali, A., Castellani, S., Di Castro, C. & Grilli, M. d -wave superconductivity near charge instabilities. *Phys. Rev. B* **54**, 16216–16225 (1996).
5. Tallon, J. L. *et al.* Critical doping in overdoped high- T_c superconductors: a quantum critical point? *Phys. Status Solidi B* **215**, 531–540 (1999).
6. Kivelson, S. A., Fradkin, E. & Emery, V. J. Electronic liquid-crystal phases of a doped Mott insulator. *Nature* **393**, 550–553 (1998).
7. Varma, C. M. Pseudogap phase and the quantum-critical point in copper-oxide metals. *Phys. Rev. Lett.* **83**, 3538–3541 (1999).
8. Chakravarty, S., Laughlin, R. B., Morr, D. K. & Nayak, C. Hidden order in the cuprates. *Phys. Rev. B* **63**, 094503 (2001).
9. Sachdev, S. Colloquium: Order and quantum phase transitions in the cuprate superconductors. *Rev. Mod. Phys.* **75**, 913–932 (2003).
10. Balakirev, F. F. *et al.* Signature of optimal doping in Hall-effect measurements on a high-temperature superconductor. *Nature* **424**, 912–915 (2003).
11. Jansen, L. & Block, R. On the relation between (maximum) critical temperature and c -axis layered structure in cuprates. I. Evaluation of existing analyses. *Physica A* **289**, 165–177 (2001).
12. Anderson, P. W. c -axis electrostatics as evidence for the interlayer theory of high-temperature superconductivity. *Science* **279**, 1196–1198 (1998).
13. Moler, K. A., Kirtley, J. R., Hinks, D. G., Li, T. W. & Ming, X. Images of interlayer Josephson vortices in $\text{TiBa}_2\text{CuO}_{6+\delta}$. *Science* **279**, 1193–1196 (1998).
14. Chakravarty, S., Kee, H.-Y. & Abrahams, E. Frustrated kinetic energy, the optical sum rule, and the mechanism of superconductivity. *Phys. Rev. Lett.* **82**, 2366–2369 (1999).
15. Boris, A. V. *et al.* Josephson plasma resonance and phonon anomalies in trilayer $\text{Bi}_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_{10}$. *Phys. Rev. Lett.* **89**, 277001 (2002).
16. Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62**, 61–122 (1999).
17. Khlebnikov, S. Interlayer tunnelling in a non-Fermi-liquid metal. *Phys. Rev. B* **53**, R11964–R11967 (1996).
18. Arrigoni, E., Fradkin, E. & Kivelson, S. A. Mechanism of high temperature superconductivity in a striped Hubbard model. Preprint at (<http://arXiv.org/abs/cond-mat/0309572>) (2003).
19. Chakravarty, S. Do electrons change their c -axis kinetic energy upon entering the superconducting state? *Eur. Phys. J. B* **5**, 337–343 (1998).
20. Bozovic, I. *et al.* No mixing of superconductivity and antiferromagnetism in a high-temperature superconductor. *Nature* **422**, 873–875 (2003).
21. Chakravarty, S., Kee, H.-Y. & Nayak, C. Neutron scattering signature of d -density wave order in the cuprates. *Int. J. Mod. Phys. B* **15**, 2901–2909 (2001).
22. Tokunaga, Y. *et al.* Effect of carrier distribution on superconducting characteristics of the multilayered high- T_c cuprate $(\text{Cu}_{0.6}\text{Co}_{0.4})\text{Ba}_2\text{Ca}_3\text{CuO}_{12+\delta}$: Cu-63-NMR study. *Phys. Rev. B* **61**, 9707–9710 (2000).
23. Kotev, H. *et al.* Unusual magnetic and superconducting characteristics in multilayered high- T_c cuprates: ^{63}Cu NMR study. *Phys. Rev. B* **64**, 064515 (2001).
24. Emery, V. J. & Kivelson, S. A. Importance of phase fluctuations in superconductors with small superfluid density. *Nature* **374**, 434–437 (1995).
25. Carlson, E. W., Kivelson, S. A., Emery, V. J. & Manousakis, E. Classical phase fluctuations in high temperature superconductors. *Phys. Rev. Lett.* **83**, 612–615 (1999).
26. Feng, D. L. *et al.* Electronic structure of the trilayer cuprate superconductor $\text{Bi}_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_{10+\delta}$. *Phys. Rev. Lett.* **88**, 107001 (2002).
27. Sato, T. *et al.* Low energy excitation and scaling in $\text{Bi}_2\text{Sr}_2\text{Ca}_{n-1}\text{Cu}_n\text{O}_{2n+4}$ ($n = 1 - 3$): angle-resolved photoemission spectroscopy. *Phys. Rev. Lett.* **89**, 067005 (2002).
28. McElroy, K. *et al.* Relating atomic-scale electronic phenomena to wave-like quasiparticle states in superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Nature* **422**, 592–596 (2003).

29. Yamauchi, H., Karppinen, M. & Tanaka, S. Homologous series of layered cuprates. *Physica C* **263**, 146–150 (1996).
30. Kuzemskaya, I. G., Kuzemsky, A. L. & Cheglov, A. A. Superconducting properties of the family of mercuric cuprates and role of layered structure. *J. Low-Temp. Phys.* **118**, 147–152 (2000).

Acknowledgements We thank the Aspen Centre for Physics, where this collaboration was initiated, and also N. P. Armitage and J. Hoffman for discussions. We acknowledge support from the US National Science Foundation (S.C.), the Canadian Institute for Advanced Research (H.-Y.K. and K.V.), and the Alfred P. Sloan Foundation (H.-Y.K.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.C. (sudip@physics.ucla.edu).

Crystal symmetry and the reversibility of martensitic transformations

Kaushik Bhattacharya¹, Sergio Conti², Giovanni Zanzotto³ & Johannes Zimmer²

¹Division of Engineering and Applied Science, California Institute of Technology, Pasadena, California 91125, USA

²Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany

³Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università di Padova, Via Belzoni 7, 35131 Padova, Italy

Martensitic transformations are diffusionless, solid-to-solid phase transitions, and have been observed in metals, alloys, ceramics and proteins^{1,2}. They are characterized by a rapid change of crystal structure, accompanied by the development of a rich microstructure. Martensitic transformations can be irreversible, as seen in steels upon quenching¹, or they can be reversible, such as those observed in shape-memory alloys^{3,4}. In the latter case, the microstructures formed on cooling are easily manipulated by loads and disappear upon reheating. Here, using mathematical theory and numerical simulation, we explain these sharp differences in behaviour on the basis of the change in crystal symmetry during the transition. We find that a necessary condition for reversibility is that the symmetry groups of the parent and product phases be included in a common finite symmetry group. In these cases, the energy barrier to lattice-invariant shear is generically higher than that pertaining to the phase change and, consequently, transformations of this type can occur with virtually no plasticity. Irreversibility is inevitable in all other martensitic transformations, where the energy barrier to plastic deformation (via lattice-invariant shears, as in twinning or slip) is no higher than the barrier to the phase change itself. Various experimental observations confirm the importance of the symmetry of the stable states in determining the macroscopic reversibility of martensitic transformations.

Martensitic transformations have numerous technological applications, particularly in steel, where the transformation induced by quenching (fast cooling) is exploited for enhancing the alloy's strength¹. Another application is the fascinating shape-memory effect in alloys like Nitinol, used in medical and engineering devices³. Martensitic phase changes are also exploited to toughen structural ceramics⁵ such as zirconia, and are observed in biological systems such as the tail sheath of the T4 bacteriophage virus⁶. The study of these transformations has led to improved materials for actuation (ferromagnetic shape-memory alloys^{7,8} and ferroelectrics⁹) and to candidates for artificial muscles¹⁰. Finally, the rich

microstructure (distinctive patterns developed at scales ranging from a few nanometres to a few micrometres) that accompanies these transformations provides a valuable test case for the development of multi-scale modelling tools¹¹.

In some materials, such as shape-memory alloys, the martensitic phase change is almost perfectly reversible (often termed 'thermo-elastic')^{3,4}. When cooling a shape-memory alloy like Nitinol from high temperature, the transformation produces a microstructure with a (twinned) plate-like morphology. This microstructure can easily be changed by the application of loads. The transformation can be completely reversed, with the disappearance of the microstructure upon reheating and the appearance of little or no dislocation or twinning in the parent (high temperature, high symmetry) phase.

In contrast, the martensitic transformation is not reversible in materials like steel and other alloys, such as CoNi¹. In steels, the microstructure forms in a sudden burst with a lath-like morphology upon cooling, and is immobile on loading. Moreover, the transformation is irreversible and the microstructure does not disappear upon reheating. In CoNi the phase change is also irreversible, as successive microstructures form both on heating and cooling. Materials that undergo irreversible transformations are characterized by significant dislocations and twinning in the parent phase.

We provide an explanation for this difference in reversibility on the basis of the symmetry change during the transformation. We label as 'weak' the martensitic transformations in which the symmetry group of both the parent and product phase are included in a common finite symmetry group¹² (which includes symmetry breaking), and all others as 'reconstructive'¹³ (note that this is different from the usage of the term 'reconstructive' to mean 'diffusional': all the transformations considered here are diffusionless). We show through rigorous mathematical theory and numerical simulation that irreversibility is inevitable in a reconstructive phase transformation, but not in a weak one.

Figure 1 illustrates our main idea through a square-to-hexagonal reconstructive phase change in a two-dimensional crystal. Consider the square lattice shown on the left, and suppose the solid square unit cell is transformed to the solid unit cell of the hexagonal lattice (middle). By symmetry the solid and dashed unit cells shown in the middle are equivalent. If the crystal is transformed back to the square phase, the dashed hexagonal cell can go to, say, the dashed cell on the right. Crucially, the square cell on the left is then transformed to the sheared cell of the square lattice on the right. Thus, upon transforming, performing a symmetry operation, and transforming back, the crystal has undergone a lattice-invariant shear, that is, a shearing deformation that leaves the entire (ideal,

infinite) lattice invariant. This also holds for finite lattices if boundary effects can be neglected, as illustrated below by means of numerical simulations, showing dislocations and plastic deformation.

Various experimental observations confirm our predictions. Pure iron (Fe) undergoes a temperature-induced reconstructive γ - α (face-centred-cubic to body-centred-cubic, f.c.c.-to-b.c.c.) transformation. As Ni and C are added to obtain steel, the martensite becomes body-centred-tetragonal (b.c.t.) with an increasing tetragonality, so that the martensitic transformation becomes weak. Correspondingly, Maki and Tamura¹⁴ found that with increasing Ni and C the reversibility of the phase changes increased, the amount of plastic deformation decreased, and the morphology changed from lath to butterfly to lenticular to plate-like. Iron also undergoes a pressure-induced reconstructive α - ϵ (b.c.c. to hexagonal-close-packed, h.c.p.) transformation, again accompanied by twins and dislocations^{15,16}. Another set of observations concern materials undergoing the reconstructive f.c.c.-to-h.c.p. transformation. Liu *et al.*¹⁷ considered the CoNi system; they hardened the parent phase to prevent any plastic deformation, but found that the transformation was still irreversible, with both heating and cooling producing twins and plastic deformation. Similar observations were made in FeMnCrSiNi steels^{18,19}. Finally, in a block co-polymer, the (reconstructive) b.c.c.-to-hexagonal transformation was observed to be irreversible because of orientation proliferation through twinning in both phases²⁰.

In spirit, our results are related to those of Otsuka and Shimizu²¹, who discuss the effects of ordering on crystallographic reversibility of martensitic transformation in alloys. They²¹ observe that in ordered alloys the transformation path must be such that the order is not destroyed, whereas in disordered ones only the atomic positions need to be recovered. Their results are consistent with many experiments, but they also note that the f.c.c.-to-face-centred-tetragonal (f.c.t.) transformation is an 'exception', which, they argue, is a consequence of the fact that "the lattice correspondence is unique in the reverse transformation because of the so simple lattice change and lower symmetry of the f.c.t. phase". We show here that the essential difference resides in the crystal symmetry, rather than in order and disorder. This provides a general framework which includes both their²¹ theory and exception, and also extends to other situations.

The common explanation (in ref. 14, for example) for the irreversibility in Fe (or low-Ni or -C steels) is based on the volume change that accompanies the transformation. The idea is that a partially transformed region causes stress and plastic deformation. However, a system like the styrene block co-polymer of ref. 20 should be largely unaffected by any volume change. Our explanation is independent of such a volume effect. It is possible, however, that both mechanisms contribute to the irreversibility in Fe.

As our result is based only on the symmetry of the energetic landscape, it is independent of material parameters, such as elastic moduli. We show that the barrier to lattice-invariant shears in reconstructive transformations is as high as to the phase transition itself. A remarkable implication is that reconstructive transformations are accompanied by plastic deformation through dislocations and twinning in the parent phase, making these phase changes irreversible. In contrast, weak transformations have the potential to be reversible, because the energy barriers to lattice-invariant shears and to the phase transition are independent of each other. Our numerical simulations indicate that weak transformations are indeed generically reversible. Yet the two energy barriers might happen to be comparable in particular materials undergoing a weak transformation. This, for example, is the case in Ni-50Ti, where plastic deformation masks the reversibility. However, (precipitate) hardening this material makes these barriers different, thereby revealing the reversible character of the underlying weak

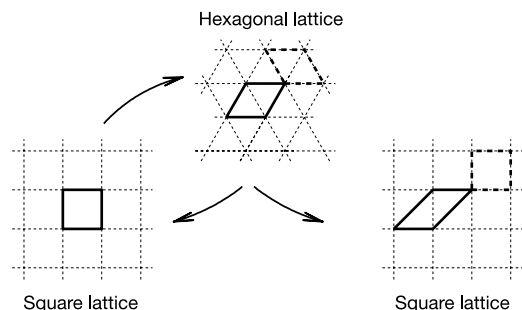


Figure 1 A lattice-invariant shear can be generated by a forward and reverse square-to-hexagonal phase transformation. The transformation takes the solid square on the left to the solid rhombus in the middle. The hexagonal symmetry implies the equivalence of the solid rhombic cell to the dashed rhombic one. The reverse transformation takes the latter to the dashed square on the right. In the process, the original solid square on the left has sheared by a lattice-invariant shear to the solid parallelogram on the right.

transformation³. In contrast, hardening does not rescue the irreversibility of CoNi¹⁷ or FeMnCrSiNi^{18,19}, which undergo reconstructive transformations. In summary, a transformation must be weak to be reversible.

Another consequence of our remarks is that materials like FeNi should be comparatively softer at compositions that are closer to a reconstructive transformation. This effect, however, may be obscured by the fact that the very softness of the ideal lattice produces plastic deformation and the entanglement of dislocations in the crystal, leading eventually to hardening. None of these phenomena are present in crystals undergoing weak martensitic transformations.

We now present the general theory followed by a concrete example, and numerical simulations. We state the theory for the simple case of Bravais lattices, which can readily be extended to crystals whose translational symmetry is continuous across the phase change. A Bravais lattice $\mathcal{L}(e_i)$ is given by the linear combinations, with integral coefficients, of three independent vectors $\{e_1, e_2, e_3\}$ forming the lattice basis. Another basis $\{f_1, f_2, f_3\}$ generates the same lattice, $\mathcal{L}(e_i) = \mathcal{L}(f_i)$, if and only if $f_i = \sum_{j=1}^3 m_{ij} e_j$ with a matrix m belonging to the group^{12,22,23}:

$$\mathcal{G} := \{m : m_{ij} \text{ integers and } \det(m) = \pm 1\} \quad (1)$$

which is the global symmetry group of Bravais lattices. Applied to any given lattice, $\mathcal{G}$ consists of rotations, reflections, lattice-invariant shears and combinations thereof; the restriction to rotations and reflections gives the lattice group (a matrix representation of the point group) of that lattice.

The free-energy density Φ of the crystal is a function of the lattice basis at a fixed temperature. It is invariant under the global symmetry group $\mathcal{G}$, as it cannot distinguish among bases generating the same lattice^{11,23,24}:

$$\Phi(e_i) = \Phi(m_{ij} e_j) \text{ for every } m \in \mathcal{G} \quad (2)$$

This global framework takes all possible deformations into account, including large shearing distortions. From equation (2), the energy landscape of the crystal has infinitely many wells, which are not contained in any bounded region in strain space (see Fig. 2).

For weak martensitic transformations the invariance of the energy (equation (2)) can be limited to a finite subgroup of $\mathcal{G}$ owing to the group-subgroup relation; correspondingly, the domain of the energy can be restricted to a neighbourhood of the reference configuration. This neighbourhood does not contain any lattice-invariant shears and only contains a finite number of energy wells. Such domains are called Ericksen-Pitteri neighbourhoods (EPNs)^{24,25}; see Fig. 2. Therefore, the classical framework of Landau theory and nonlinear elasticity^{11,23,26} applies in these cases.

For reconstructive martensitic transformations, on the other hand, the symmetry decreases along the transformation path, but increases again at the final state, and there is no reference configuration whose lattice group contains those of the two given phases.

We establish the following mathematical fact: the lattice groups of the two phases in any reconstructive transformation necessarily generate unbounded shear-like distortions. Consequently, the barrier between the infinitely many energy wells of the crystal is, at most, equal to that of the underlying phase change. This implies that the material cannot resist certain arbitrarily large deformations, which makes the reconstructive transformation necessarily non-thermoelastic and irreversible, owing to the creation of defects in the lattice. At the same time, no reduction to a finite number of wells in a bounded region is possible (no EPN can be extracted) and the classical approach of the Landau type is not applicable. (Some authors have extended the Landau framework to encompass reconstructive transformations by introducing a “transcendental order parameter”¹³, which gives a partial description of the lattice periodicity characterized by the global group $\mathcal{G}$.)

We show in the Methods section that in a reconstructive transformation an unbounded element of $\mathcal{G}$ is always generated. Precisely, we necessarily obtain an element with an infinite period (that is whose powers can become arbitrarily large), akin to slip and twinning in the parent phase. We note further that the most common reconstructive transformations involve phases with maximal point symmetry: the primitive-cubic, the f.c.c., the b.c.c., or the hexagonal subgroups of $\mathcal{G}$. For instance, consider an f.c.c. lattice. For some transformations to b.c.c., such as the Bain stretch considered below, the symmetry groups of the two lattices generate the entire group $\mathcal{G}$, with its full set of lattice-invariant shears. The same happens for any transformation between phases with maximal symmetry.

The impossibility of restricting the symmetry to a finite subgroup of $\mathcal{G}$, and the energy domain to a suitable EPN, has dramatic implications for the variational treatment of reconstructive phase transformations. Indeed, if we assume that the deformation at each time is determined by minimizing the free-energy subject to external forces and boundary conditions, the invariance of the energy under the whole group $\mathcal{G}$ implies that the solid cannot resist any shear²⁷. In practice, dynamics and defects, including dislocations, moderate this phenomenon, which we revisit through our numerical simulations.

We now focus on a concrete case study, demonstrating that the

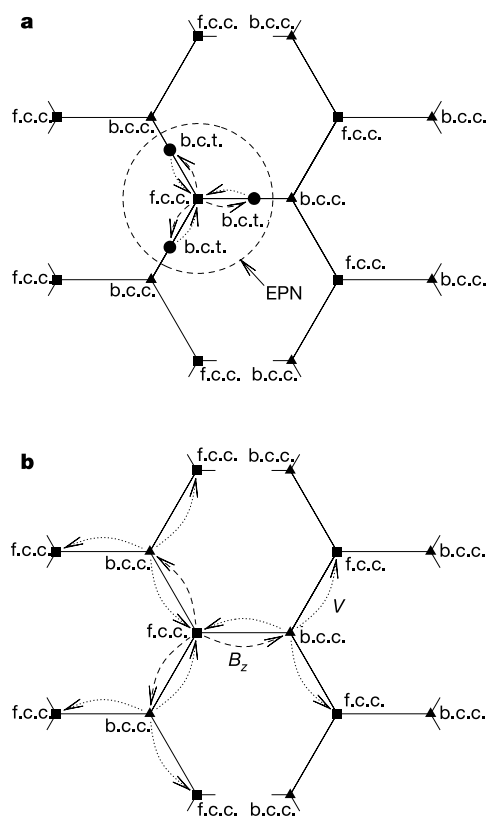


Figure 2 Schematic representation of weak versus reconstructive transformations in the space of lattices. **a**, Weak. A uniaxial deformation (dashed arrows) of the f.c.c. lattice (represented by squares) can give three equivalent b.c.t. lattices (circles). The reverse transformation (dotted arrows) returns to the original f.c.c. configuration. All the transformation strains are confined within a single EPN, such as the dashed circle. **b**, Reconstructive f.c.c.-to-b.c.c.. The transformation B_2 leads from an f.c.c. lattice to three equivalent b.c.c. lattices, and the reverse transformation (such as V) from each b.c.c. lattice can proceed to three distinct but equivalent f.c.c. lattices, and so on. No EPN can be singled out.

composition of an f.c.c.-to-b.c.c. (forward) transformation and a b.c.c.-to-f.c.c. (reverse) transformation can in fact result in a lattice-invariant shear. We start from an f.c.c. lattice aligned with the coordinate axes, and subject it to a uniaxial stretch U_z along the z axis:

$$U_z^{(\lambda)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/\lambda \end{pmatrix},$$

neglecting volumetric changes. For λ between 1 and $\sqrt{2}$, the product lattice has tetragonal symmetry, and we have a weak transformation. When λ is equal to $\sqrt{2}$, we obtain the Bain stretch $U_z^{(\sqrt{2})} = B_z$ which produces the b.c.c. lattice in a reconstructive transformation: see Fig. 3. Now consider the reverse b.c.c.-to-f.c.c. transformation: it can occur in three possible ways, owing to the symmetry of the b.c.c. phase. A crucial point is that the four-fold symmetry axes of the b.c.c. lattice are different from those of the initial f.c.c. crystal. One of the three possibilities for the reverse stretch corresponds to B_z^{-1} , and leads back to the original lattice. The other two are stretches along the $(1, \pm 1, 0)$ directions (still using the original coordinate system). For instance, let us consider the reverse stretch V along the $(1, 1, 0)_{\text{f.c.c.}}$ direction, which can be expressed as

$$V = QB_z^{-1}Q^T$$

where Q is the 90-degree rotation with axis $(1, -1, 0)$, which belongs to the point group of the b.c.c. phase. After applying V , the final lattice is again f.c.c.; however, the total transformation gives:

$$VB_z = QB_z^{-1}Q^TB_z = RS$$

where R is a rotation and S is a shear on a $\{111\}_{\text{f.c.c.}}$ plane along a

$\langle 112 \rangle_{\text{f.c.c.}}$ direction. We can check that S brings the original f.c.c. basis to a $\mathcal{G}$ -equivalent one, that is, the transformation VB_z restores the f.c.c. lattice to itself up to the inessential rigid-body rotation R (see Fig. 3). Successive iterations of VB_z then generate larger and larger lattice-invariant deformations, and similar strategies can generate the entire group $\mathcal{G}$. The barrier between the $\mathcal{G}$ -related sheared configurations originated in this way is exactly equal to the barrier pertaining to the phase transformation. If only part of the crystal undergoes the VB_z transformation, while the rest goes back to the original state, a Shockley partial dislocation with Burgers vector $\frac{1}{6}\langle 112 \rangle$ is formed²⁸. The double transformation $(VB_z)^2$ then gives the Burgers vector $\frac{1}{2}\langle 110 \rangle$. Both kinds are typical of f.c.c. crystals. When starting from a b.c.c. crystal, this same mechanism generates dislocations with Burgers vector $\langle 111 \rangle_{\text{b.c.c.}}$ on $\{112\}_{\text{b.c.c.}}$ planes, which are among the most common ones in b.c.c. lattices.

We also illustrate the phenomena that are discussed above for finite domains in the presence of boundary conditions, with numerical simulations done for simplicity for a square-to-hexagonal transformation in two dimensions. We consider a 50×50 grid of atoms interacting with a three-body nearest-neighbour potential²⁹, which produces for the crystal a two-dimensional version of the $\mathcal{G}$ -invariant energy in equation (2).

The simulation of a shearing experiment of a crystal close to a reconstructive phase transformation is shown in Fig. 4b. As the left and the right boundaries are progressively displaced, dislocations form in the lattice, which lead to irreversible plastic deformations and defects in the crystal.

In contrast, Fig. 4d shows the same simulation for a crystal close to a square-to-rhombic symmetry-breaking (weak) transformation. In this case, no dislocations arise in the lattice. Instead, typical layered martensitic twins mixing with the higher-symmetry parent

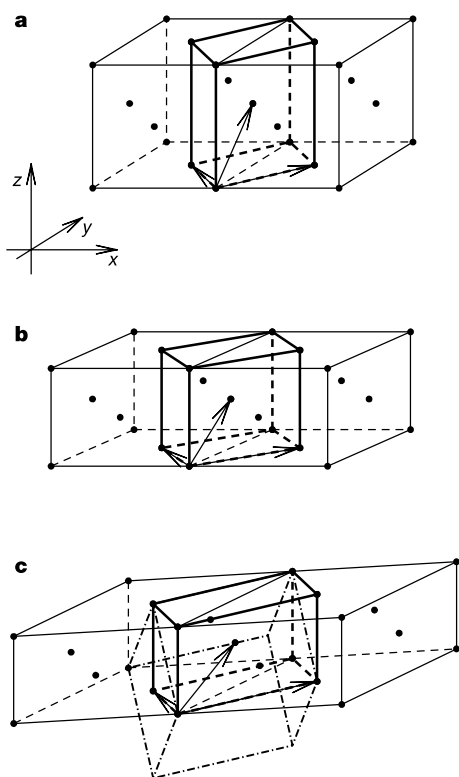


Figure 3 Shear generated for an f.c.c.-to-b.c.c. transformation in an f.c.c.-b.c.c.-f.c.c. cycle. **a**, An f.c.c. lattice, with a b.c.c. cell highlighted. **b**, The uniaxial Bain stretch B_z transforms the b.c.c. cell to the b.c.c. one. **c**, The inverse Bain stretch $V = QB_z^{-1}Q^T$ transforms the crystal back to f.c.c. (dash-dotted bold f.c.c. cell highlighted), but sheared relative to the original. The arrows show a basis undergoing the lattice-invariant shear.

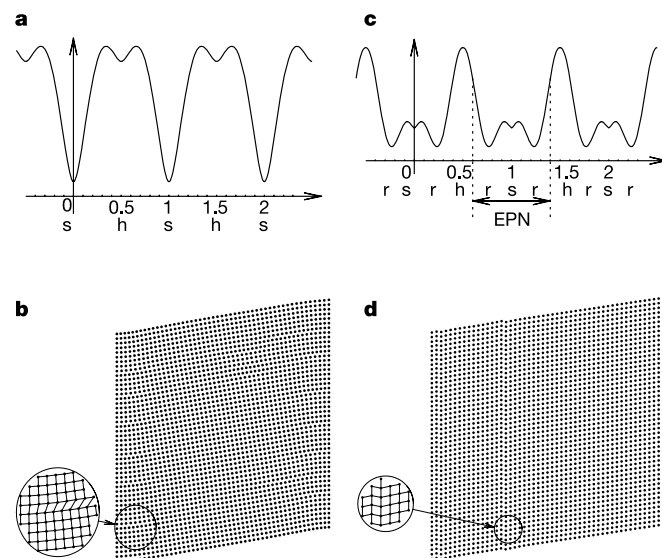


Figure 4 Reconstructive transformations generate dislocations, weak ones do not. **a, b**, Reconstructive transformation. **c, d**, Weak transformation. **a**, Section of the energy profile of a crystal close to a square-to-hexagonal (s-h) transformation, along a s-h-s line. The energy has the full invariance (2), with minima s and h plotted respectively at positions 0, 1, and 0.5, 1.5, and so on. **b**, An incremental shear test with boundary conditions on the left and right side results in dislocations. **c**, Section of the generic energy profile of a crystal close to a square-to-rhombic (s-r) transformation, along a s-r-h line. The square states s in 0, 1, and so on are metastable, and additional minima are present at intermediate rhombic configurations r, with hexagonal maxima at 0.5, 1.5, and so on. **d**, The same incremental shear test results in no dislocations, but only reversible transformation twins, because starting from the square minimum at 0, the 'faraway' square at 1 is not reachable by overcoming a small energy barrier.

phase can be observed, which accommodate the imposed boundary condition in a reversible way (this is the origin of the memory effect). □

Methods

To prove that in a reconstructive transformation an aperiodic element of $\mathcal{G}$ (called GL(3, $\mathbb{Z}$) in algebra) is generated, we first note that any lattice group is finite, and conversely any finite subgroup of $\mathcal{G}$ is included in the lattice group of some lattice (see proposition 3.5 in ref. 23). Thus a transformation is weak if and only if the lattice groups of the two crystal phases generate a finite group. Therefore a reconstructive transformation produces an infinite subgroup of $\mathcal{G}$ with a finite number of generators. Such a group necessarily contains an element with no finite period as a consequence of the Burnside–Schur theorem on periodic groups.

We finally establish that for any pair of Bravais lattices with maximal point symmetry there are reconstructive transformations that generate the entire group $\mathcal{G}$. Indeed, it is readily verified that for suitable pairs of subgroups in $\mathcal{G}$ belonging to the four arithmetic classes with maximal point symmetry one can produce all the generators of $\mathcal{G}$, that is, a suitable reflection, permutation and simple shear³⁰.

Received 2 September 2003; accepted 28 January 2004; doi:10.1038/nature02378.

- Olson, G. B. & Owen, W. (eds) *Martensite* (ASM International, Materials Park, OH, 1992).
- Sajje, E. K. H. *Phase Transitions in Ferroelastic and Co-elastic Crystals* (Cambridge Univ. Press, Cambridge, 1993).
- Otsuka, K. & Wayman, C. M. *Shape Memory Materials* (Cambridge Univ. Press, Cambridge, 1998).
- Barrett, C. S. & Massalski, T. B. *Structure of Metals* 3rd edn (Pergamon, Oxford, 1987).
- Bocanegra-Bernal, M. H. & De la Torre, S. D. Phase transitions in zirconium dioxide and related materials for high performance engineering ceramics. *J. Mater. Sci.* **37**, 4947–4971 (2002).
- Olson, G. B. & Hartman, H. Martensite and life—displacive transformations as biological processes. *J. Physique* **43**(C4), 855–865 (1982).
- James, R. D. & Wuttig, M. Magnetostriction of martensite. *Phil. Mag. A* **77**, 1273–1299 (1998).
- Sozinov, A., Likhachev, A. A., Lanska, N. & Ullakko, K. Giant magnetic-field-induced strain in NiMnGa seven-layered martensitic phase. *Appl. Phys. Lett.* **80**, 1746–1748 (2002).
- Shu, Y. C. & Bhattacharya, K. Domain patterns and macroscopic behaviour of ferroelectric materials. *Phil. Mag. B* **81**, 2021–2054 (2001).
- de Gennes, P.-G. & Okumura, K. Phase transitions of nematic rubbers. *Europhys. Lett.* **63**, 76–82 (2003).
- Bhattacharya, K. *Microstructure of Martensite: Why it Forms and How it Gives Rise to the Shape-Memory Effect* (Oxford Univ. Press, Oxford, 2003).
- Erickson, J. L. Weak martensitic transformations in Bravais lattices. *Arch. Ration. Mech. Anal.* **107**, 23–36 (1989).
- Tolédano, P. & Dmitriev, V. *Reconstructive Phase Transitions* (World Scientific, Singapore, 1996).
- Maki, T. & Tamura, I. in *Proc. Int. Conf. on Martensitic Transformations* 963–970 (The Japan Institute of Metals, Nara, 1986).
- Barker, L. M. & Hollenbach, R. E. Shock wave study of the $\alpha \leftrightarrow \epsilon$ phase transition in iron. *J. Appl. Phys.* **45**, 4872–4887 (1974).
- Kadau, K., Germann, T. C., Lomdahl, P. S. & Holian, B. L. Microscopic view of structural phase transitions induced by shock waves. *Science* **296**, 1681–1684 (2002).
- Liu, Y. et al. Thermomechanical behaviour of fcc $\leftrightarrow$ hcp martensitic transformation in CoNi. *J. Alloys Compd.* (in the press).
- Matsumoto, S., Sato, A. & Mori, T. Formation of h.c.p. and f.c.c. twins in an Fe–Mn–Cr–Si–Ni alloy. *Acta Metall. Mater.* **42**, 1207–1213 (1994).
- Yang, J. H. & Wayman, C. M. On secondary variants formed at intersections of ϵ martensite variants. *Acta Metall. Mater.* **40**, 2011–2023 (1992).
- Lee, H. H. et al. Orientational proliferation and successive twinning from thermoreversible hexagonal-body-centered cubic transitions. *Macromolecules* **35**, 785–794 (2002).
- Otsuka, K. & Shimizu, K. On the crystallographic reversibility of martensitic transformations. *Scripta Met.* **11**, 757–760 (1977).
- Schwarzenberger, R. L. E. Classification of crystal lattices. *Proc. Camb. Phil. Soc.* **72**, 325–349 (1972).
- Pitteri, M. & Zanzotto, G. *Continuum Models for Phase Transitions and Twinning in Crystals* (Chapman & Hall/CRC, Boca Raton, 2002).
- Erickson, J. L. Some phase transitions in crystals. *Arch. Ration. Mech. Anal.* **73**, 99–124 (1980).
- Pitteri, M. Reconciliation of local and global symmetries of crystals. *J. Elast.* **14**, 175–190 (1984).
- Ball, J. M. & James, R. D. Proposed experimental tests of a theory of fine microstructure and the two-well problem. *Phil. Trans. R. Soc. Lond. A* **338**, 389–450 (1992).
- Fonseca, I. Variational methods for elastic crystals. *Arch. Ration. Mech. Anal.* **97**, 189–220 (1987).
- Hull, D. & Bacon, D. J. *Introduction to Dislocations* 3rd edn (Pergamon, Oxford, 1984).
- Conti, S. & Zanzotto, G. A variational model for reconstructive phase transformations in crystals, and their relation to dislocations and plasticity. *Arch. Ration. Mech. Anal.* (in the press).
- Hua, L. K. & Reiner, I. On the generators of the symplectic modular group. *Trans. Am. Math. Soc.* **65**, 415–426 (1949).

Acknowledgements This work was largely carried out when J.Z. held a position at the California Institute of Technology. The work of S.C. and J.Z. was partially supported by the Deutsche Forschungsgemeinschaft. G.Z. acknowledges the partial support of the Italian MIUR (CoFin Modelli Matematici per i Materiali). S.C. and G.Z. acknowledge the partial support of the IV Framework Programme of the EU. K.B. and J.Z. acknowledge the partial financial support of the US Air Force Office of Scientific Research and the US Office of Naval Research. All authors contributed equally to this work.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.B. (bhattacha@caltech.edu).

Polar ocean stratification in a cold climate

Daniel M. Sigman¹, Samuel L. Jaccard² & Gerald H. Haug³

¹Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA

²Department of Earth Sciences, ETH, CH-8092 Zürich, Switzerland

³Geoforschungszentrum Potsdam, D-14473 Potsdam, Germany

The low-latitude ocean is strongly stratified by the warmth of its surface water. As a result, the great volume of the deep ocean has easiest access to the atmosphere through the polar surface ocean. In the modern polar ocean during the winter, the vertical distribution of temperature promotes overturning, with colder water over warmer, while the salinity distribution typically promotes stratification, with fresher water over saltier. However, the sensitivity of seawater density to temperature is reduced as temperature approaches the freezing point, with potential consequences for global ocean circulation under cold climates^{1,2}. Here we present deep-sea records of biogenic opal accumulation and sedimentary nitrogen isotopic composition from the Subarctic North Pacific Ocean and the Southern Ocean. These records indicate that vertical stratification increased in both northern and southern high latitudes 2.7 million years ago, when Northern Hemisphere glaciation intensified in association with global cooling during the late Pliocene epoch. We propose that the cooling caused this increased stratification by weakening the role of temperature in polar ocean density structure so as to reduce its opposition to the stratifying effect of the vertical salinity distribution. The shift towards stratification in the polar ocean 2.7 million years ago may have increased the quantity of carbon dioxide trapped in the abyss, amplifying the global cooling.

The Subarctic Zone in the North Pacific Ocean and the Antarctic Zone in the Southern Ocean are both characterized by year-round availability of the ‘major nutrients’ nitrate and phosphate. Nutrient-rich deep water is brought to the surface by wind-driven upwelling and density-driven overturning. Limitation of algal growth by light³ and iron⁴ prevents complete consumption of the major nutrients. The Subarctic Pacific maintains a higher degree of nutrient utilization (and thus lower surface nutrient concentrations) than does the Antarctic⁵. There are two likely causes for this difference. First, the exchange between the surface and deep ocean is reduced in the Subarctic Pacific relative to the Antarctic. This is partially due to the stronger ‘halocline’, or vertical salinity gradient, in the Subarctic Pacific⁶ (Fig. 1). Second, atmospheric deposition supplies more iron to the Subarctic Pacific than to the Antarctic, which should allow phytoplankton to consume a larger fraction of the upwelled nitrate and phosphate⁴.

Despite the differences between these two polar ocean regions, the sediments underlying them show a similar change during the global cooling from the relatively warm mid-Pliocene to the late Pliocene, when the Earth descended into the Pleistocene cycle of ice ages. In both of these regions, upon the intensification of major Northern Hemisphere glaciation 2.7 million years ago (2.7 Myr), the accumulation of biogenic opal decreased abruptly, just when the sedimentary evidence indicates an increase in Northern Hemisphere sea ice and icebergs^{7,8} (Figs 1 and 2; additional references in Fig. 1 legend). In the Antarctic, this shift has been interpreted as the result of increased sea ice cover shortening the productive season of diatoms^{8,9}, whereas the Subarctic Pacific change has been explained as the onset of permanent stratification reducing the nutrient supply to the surface⁷. Yet the similarity in the structure and timing of these changes invites a single explanation

that would apply in both hemispheres. To distinguish between the very different alternative explanations that have been posed for the 2.7-Myr decrease in opal accumulation, we have measured the $^{15}\text{N}/^{14}\text{N}$ ratio of bulk sediments as an indicator of nitrate utilization (see Methods). Oceanic phytoplankton preferentially consume ^{14}N -nitrate over ^{15}N -nitrate, resulting in lower sediment $^{15}\text{N}/^{14}\text{N}$ under more nitrate-replete conditions (that is, the more the gross physical supply of nitrate exceeds the biological assimilation of nitrate)¹⁰.

In the Subarctic Pacific record, sediment $^{15}\text{N}/^{14}\text{N}$ increases markedly at the 2.7-Myr event, simultaneously with the decrease in opal accumulation (Fig. 2a, inset) and the increase in ice-rafted debris⁷, suggesting more complete consumption of surface nitrate subsequent to the intensification of Northern Hemisphere glaciation. This result is consistent with the observation that diatom opal export in the modern Subarctic Pacific could not increase by more than 30% before silicate would be completely absent from the summertime surface layer¹¹, which indicates that the sharp drop in opal flux at 2.7 Myr cannot be attributed to a decrease in the completeness of silicate consumption. Together, the records of $^{15}\text{N}/^{14}\text{N}$ and opal accumulation rate make a strong case that the sedimentary change at 2.7 Myr records a decrease in the rate of exposure of nutrient-bearing deep water at the Subarctic Pacific surface. After the 2.7-Myr event, cycles appear, with lower opal flux and higher $^{15}\text{N}/^{14}\text{N}$ found at intervals that represent the cold phases/ice volume maxima associated with the 41-kyr (obliquity) component of the Milankovitch cycles (Fig. 2a, inset), suggestive of continued climate-induced oscillations in stratification¹².

In the Antarctic record from Ocean Drilling Program (ODP) Site 1096, over the sediment interval recording the change in opal flux to the sea floor, there is no significant change in the sediment $^{15}\text{N}/^{14}\text{N}$ (Fig. 2b). If the threefold decrease in diatom export suggested by the

opal accumulation change were driven by a decrease in the degree of nitrate consumption, the $^{15}\text{N}/^{14}\text{N}$ of sinking organic matter should have been roughly 3‰ higher before the 2.7-Myr shift¹³. As in the Subarctic Pacific, this suggests that the decrease in opal accumulation rate was not driven by a decrease in the degree to which nutrients were consumed in the surface layer. Rather, the data suggest that the supply of nutrients decreased. While silicate consumption is far from complete in parts of the Antarctic, essentially all of the silicate is consumed within $\sim 5^\circ$ latitude of the Polar Front in the regions where it has been studied¹⁴, and the 2.7-Myr decrease in opal accumulation observed at ODP Site 1096 has been noted from various sites in the Antarctic (black dots in Fig. 1, references in legend), several of which are close to the Polar Front (the dashed line in Fig. 1) and most of which are overlain by surface waters that are silicate-poor by the end of the summer⁵. Therefore, the modern surface distribution of dissolved silicate suggests strongly that the decrease in opal accumulation at 2.7 Myr was not solely due to a decrease in the opal:carbon ratio of export production¹⁵, as this change alone would require that, after 2.7 Myr, surface waters became extremely enriched in silicate. Although the modern ocean may not be an ideal reflection of mean post-2.7 Myr conditions, interglacials similar to the current one should be reflected in the post-2.7 Myr interval of the sediment records in Fig. 1. Thus, as in the Subarctic Pacific, the $^{15}\text{N}/^{14}\text{N}$ and opal data together indicate that the nutrient supply to the surface was much greater before the late Pliocene cooling event, suggestive of a subsequent shift towards stratification.

That the apparent degree of nitrate utilization increased at 2.7 Myr in the Subarctic Pacific but remained constant in the Antarctic is consistent with the iron budget in these different regions. In the Subarctic Pacific, where a significant fraction of the iron supply comes from the atmosphere⁶, algal productivity

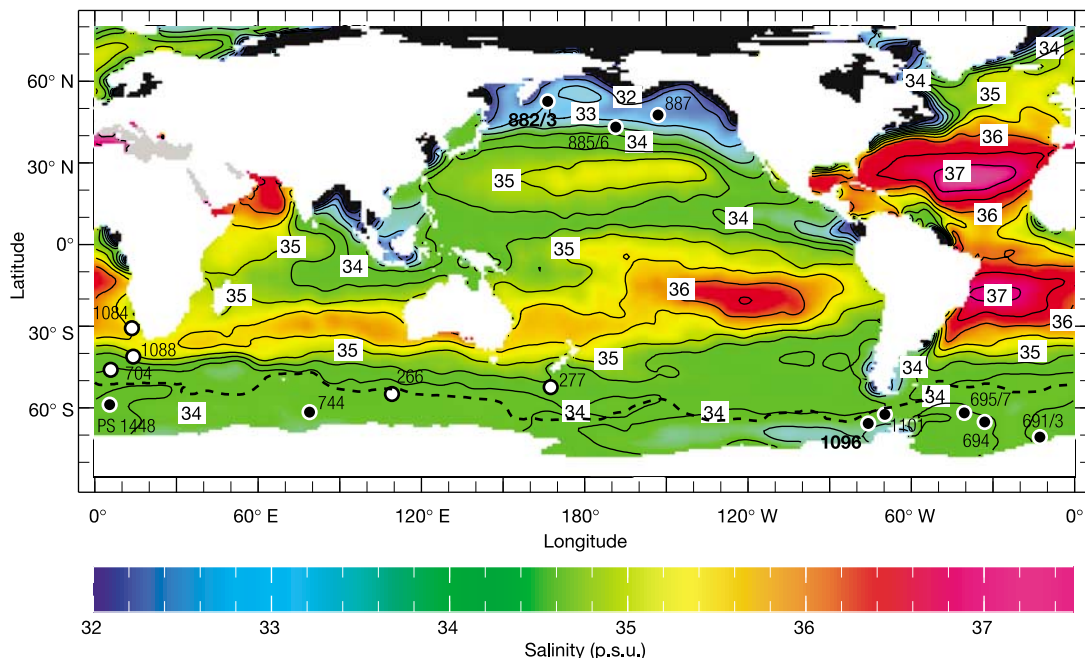


Figure 1 Map of mean annual surface salinity⁵ illustrating the predominance of low-salinity surface waters at high latitudes, including the Subarctic North Pacific and the Antarctic. The Antarctic Polar Front, the northern boundary of the Antarctic, is indicated by the dashed line in the Southern Ocean²⁶. Filled circles indicate ODP sites where biogenic opal accumulation decreased during the late Pliocene (the sediment cores shown in Fig. 2 are in bold), whereas open circles show locations where opal accumulation increased

during roughly the same time interval^{7–9,25,27–30}. The broad distribution of the opal decrease within both the Subarctic North Pacific and the Antarctic argues against lateral sediment transport as the cause of the opal accumulation decrease. The increase in opal accumulation at lower latitudes may have been driven by an increase in the oceanic silicate concentration resulting from the drop in opal burial in the polar regions.

would not be strictly limited by the amount of dissolved iron imported from deep water, so that the degree of nitrate consumption should increase in response to stratification. In contrast, in the Antarctic, where nearly all of the iron supply is from upwelled deep water, productivity should decrease as deep water supply decreases, so that the fraction of the nitrate supply that is consumed should remain constant in the face of stratification¹⁶. Moreover, the lack of any evidence for iron enrichment in the Antarctic provides an additional argument against a decrease in the opal:carbon ratio of export production at the 2.7-Myr transition.

These data fit into a growing body of evidence for a pervasive link between cold climates and polar ocean stratification during the late Cenozoic. Stronger stratification has been proposed for the Antarctic surface during late Pleistocene glacial periods^{17,18} and for the last glacial Okhotsk Sea and western Subarctic North Pacific¹². The well-studied reduction in deep water formation in the high-latitude North Atlantic during ice ages¹⁹ is, in essence, a shift towards stratification in that region. While efforts to explain the stratification of the glacial North Atlantic have focused on changing freshwater inputs²⁰, the strong correlation of reduced deep water formation with cold periods is most suggestive of cooling as the direct cause of stratification.

The vertical stability of the polar ocean is controlled by the typically opposing effects of temperature (destabilizing) and salinity (stabilizing) (Fig. 3a), but density is less dependent on temperature at lower temperatures. As a result, as the mean temperature of the ocean decreases, a given temperature difference between the surface and deep water would cause a smaller density

difference. Thus, homogeneous cooling of the polar ocean water column would cause the vertical temperature gradient to be less important in setting the density gradient, rendering the existing salinity gradient more capable of maintaining year-round polar ocean stratification. Models have shown that stratification of the North Atlantic during glacial times could be the result of global ocean cooling, given the nonlinear sensitivity of seawater density to temperature^{1,21}.

The importance of global ocean temperature to stratification in other polar regions can be illustrated by recalculating the density profile of the wintertime Antarctic under conditions of higher and lower temperature than today (Fig. 3). If the entire water column is warmed by 3 °C, the modern surface-to-subsurface density difference is removed, such that there would be no barrier to overturning (Fig. 3b, red crosses compared to bold black 'plus' symbols). On the other hand, cooling the entire water column by 2 °C causes an approximate doubling of the modern surface-to-subsurface density difference (Fig. 3c, blue crosses compared to bold black 'plus' symbols). Such a cooling is actually impossible, as the wintertime surface is already near freezing (Fig. 3a). If the entire water column is simply cooled to the freezing point, the vertical density difference is made even stronger (Fig. 3c, red crosses compared to bold black 'plus' symbols), such that a halving of the modern surface-to-deep salinity difference in the wintertime Antarctic would be required to completely counteract its effect (Fig. 3c, green crosses and dashed line). Changes in polar stratification through this mechanism would not develop until the entire ocean water column had changed temperature. Thus, this response would not be evident in model

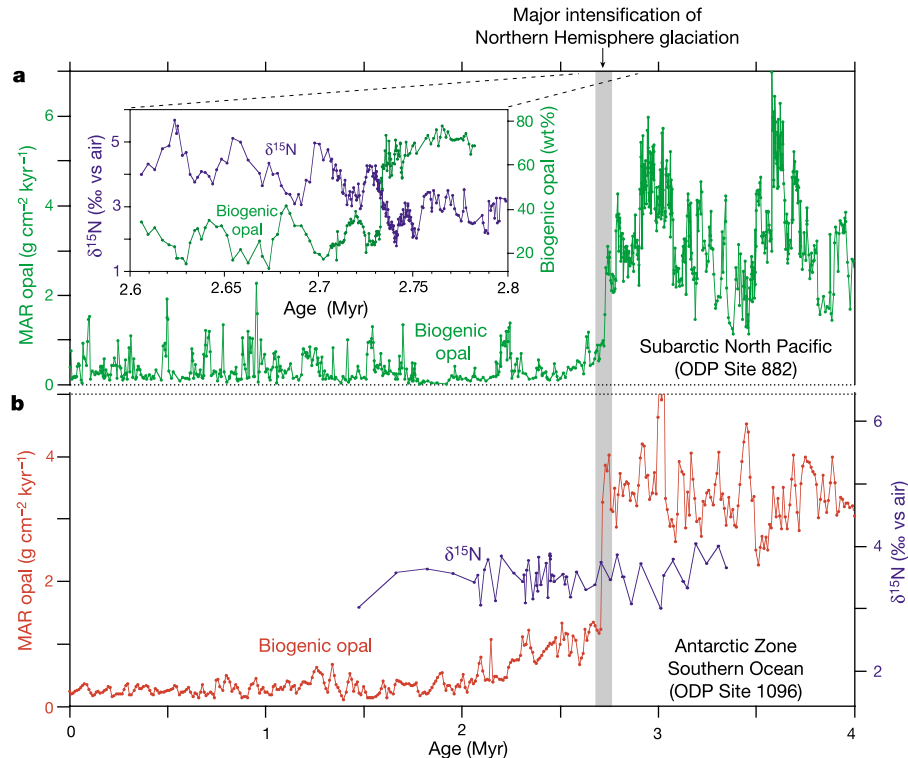


Figure 2 Subarctic North Pacific (ODP Site 882) and Southern Ocean (ODP Site 1096)⁸ palaeoceanographic time series. Shown are biogenic opal accumulation over the past 4 Myr and sediment $^{15}\text{N}/^{14}\text{N}$ across the late Pliocene opal transition. **a**, Mass accumulation rate (MAR) of biogenic opal in the Subarctic North Pacific (green) decreases at the onset of major Northern Hemisphere glaciation at 2.7 Myr (more specifically, 2.73 Myr; ref. 7). Inset, fluctuations in biogenic opal percentage (green) and $\delta^{15}\text{N}$ (blue) during the time interval 2.6–2.73 Myr reveal a cycle consistent with the Milankovitch

obliquity frequency (high sediment $\delta^{15}\text{N}$ correlates with abundance peaks in ice-rafted debris⁷ (data not shown); $\delta^{15}\text{N} = ((^{15}\text{N}/^{14}\text{N})_{\text{sample}} / (^{15}\text{N}/^{14}\text{N})_{\text{reference}} - 1) \times 1,000\text{‰}$). **b**, Mass accumulation rate of biogenic opal in the Antarctic zone of the Southern Ocean (ODP Site 1096, red)⁸ shows a decrease that is similar in structure and timing to that observed in the Subarctic North Pacific. Sediment $\delta^{15}\text{N}$ (blue) across the opal accumulation transition shows no systematic or significant change, varying between 3 and 4‰.

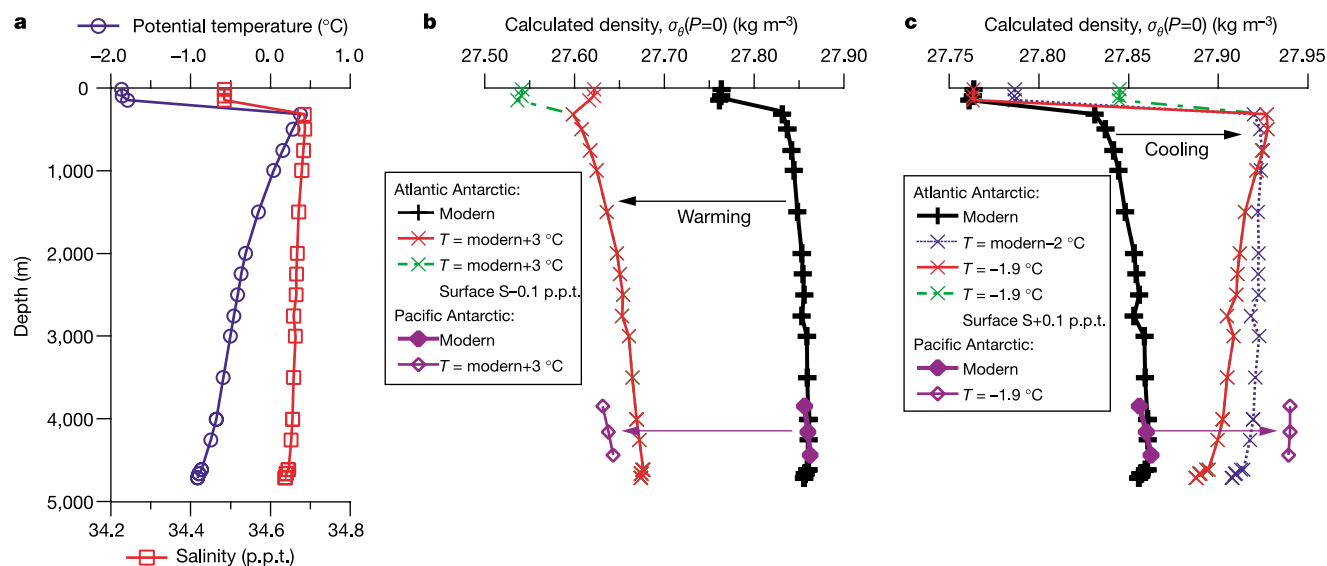


Figure 3 Density as a function of depth in the modern wintertime Antarctic and changes in this density structure for uniform changes in seawater temperature. Potential temperature (**a**, open blue circles), salinity (**a**, open red squares), and calculated potential density (**b** and **c**, bold black 'plus' symbols) for a wintertime profile in the Antarctic (WOCE SR04 station 163, Atlantic sector, 65.7° S, 38.8° W, 23 September 1989), and recalculation of potential density after several simple hypothetical changes in water column temperature and salinity (**b** and **c**, crosses). Wintertime data are used because overturning is most likely during that season, and this region is among the least vertically stable in the open Antarctic. The recalculated potential density profiles are for: (1) warming the entire water column by 3 °C (**b**, red crosses and solid line); (2) warming the entire water column by 3 °C

and decreasing the salinity of the surface mixed layer by 0.1 p.p.t. (**b**, green crosses and dot-dashed line); (3) cooling the entire water column by 2 °C (**c**, blue crosses and dotted line; this would cool the shallowest and deepest waters past the freezing point); (4) cooling the entire water column to near the freezing point, -1.9 °C (**c**, red crosses and solid line); and (5) cooling the entire water column to -1.9 °C and increasing the salinity of the surface mixed layer by 0.1 p.p.t. (**c**, green crosses and dot-dashed line). Density and recalculated density after warming (**b**) and cooling (**c**) are also plotted for deep waters in the Pacific sector of the Antarctic near ODP Site 1096 (purple diamonds, filled and open, respectively; WOCE S04P station 702, 67° S, 87° W, 28 February 1992), which are warmer and saltier than deep waters in the Atlantic sector of the Antarctic.

experiments that are run for less than the ventilation time of the deep ocean²².

One important caveat to our proposal is that, as the low latitudes cool, the atmosphere may transport less fresh water to the polar ocean, which could work to reduce the polar salinity gradient (Fig. 3, green crosses). Moreover, other physical mechanisms for a link between cooling and stratification cannot be ruled out. The westerly winds could shift equatorward and/or weaken as climate cools, leading to a decrease in wind-driven upwelling in the Antarctic²³. The sea ice cycle redistributes fresh water within the polar regions, determining where sinking and stratification occurs in the modern polar ocean, and there is evidence for the production, during the last ice age, of extremely salty bottom water in the Antarctic owing to the formation of sea ice², the melting of which may have stratified the open Antarctic. However, it remains unclear whether these other mechanisms could explain the stratification of both the Antarctic and Subarctic North Pacific at 2.7 Myr.

If a polar region migrates towards stratification, positive feedbacks reinforce this change. An increase in polar ocean stratification would increase the residence time of water in the polar surface by decreasing exchange with the subsurface, allowing the surface to accumulate the fresh water being deposited from the atmosphere and thus causing further stratification. As a result, while gradual ocean cooling would motivate a gradual shift towards stratification, the positive feedback associated with freshwater accumulation may cause the onset of stratification to be abrupt, as is observed for the 2.7-Myr transition.

A positive feedback of greater global significance may operate through atmospheric carbon dioxide. The export of organic carbon out of the low-latitude surface ocean works to sequester carbon dioxide in the ocean interior, while the exposure of deep water at the polar ocean surface allows this carbon dioxide to escape back to

the atmosphere if not all of the major nutrients are consumed there. The slower the ocean is ventilated at a high-nutrient polar surface region, the more the low-latitude 'biological pump' is able to drive the accumulation of surface-extracted carbon dioxide in the ocean interior²⁴. By preventing the polar release of deeply sequestered carbon dioxide, stratification in both the Antarctic and the Subarctic Pacific may have amplified the climate cooling associated with the onset of Northern Hemisphere glaciation, while overturning of the polar ocean may have played a role in the poorly understood warmings of the late Oligocene and middle Miocene (see Supplementary Information). □

Methods

Ocean Drilling Program Sites 882 and 1096 are in the northwest Subarctic Pacific (50° 21' N, 167° 35' E; water depth 3,244 m) and eastern Pacific sector of the Antarctic (67° 34' S, 76° 58' W, water depth 3,152 m), respectively. Methods for the opal mass accumulation records of ODP Sites 882 and 1096 are described in refs 7 and 8, respectively, while the biogenic opal percentage record for ODP Site 882 in the inset of Fig. 2 was determined by alkaline extraction of silica²⁵, with replication indicating a reproducibility of $\pm 3\%$.

Nitrogen isotope ratios were determined on freeze-dried and homogenized samples by on-line combustion using a ThermoQuest NC2500 elemental analyser coupled to a ThermoFinnigan DeltaPlus mass spectrometer via a ThermoFinnigan ConFlo III at the University of British Columbia, Vancouver. Analytical precision is $\pm 0.3\text{‰}$ (1 σ), and the values are reported relative to air N₂.

The astronomically calibrated stratigraphy for ODP Site 882 is based on magnetostratigraphy and fine tuning of GRAPE density oscillations in the orbital precession band to the summer insolation at 65° N (ref. 7). The age model for ODP Site 1096 is based on magnetostratigraphy⁸. At ODP Site 1096, the sharpest decrease in biogenic opal percentage occurs roughly 12 m below the Matuyama/Gauss magnetic reversal at 2.58 Myr, suggesting that the bulk accumulation rate change is at roughly 2.7 Myr, consistent with the established significance of 2.73 Myr for climate and sediment records in general. Accordingly, a subtle change has been applied to the original age model.

Received 4 June 2003; accepted 19 January 2004; doi:10.1038/nature02357.

1. Winton, M. The effect of cold climate upon North Atlantic Deep Water formation in a simple ocean-atmosphere model. *J. Clim.* **10**, 37–51 (1997).

2. Adkins, J. F., McIntyre, K. & Schrag, D. P. The salinity, temperature, and $\delta^{18}\text{O}$ of the glacial deep ocean. *Science* **298**, 1769–1773 (2002).
3. Mitchell, B. G., Brody, E. A., Holm-Hansen, O., McClain, C. & Bishop, J. Light limitation of phytoplankton biomass and macronutrient utilization in the Southern Ocean. *Limnol. Oceanogr.* **36**, 1662–1677 (1991).
4. Martin, J. H. & Gordon, R. M. Northeast Pacific iron distributions in relation to phytoplankton productivity. *Deep-Sea Res.* **35**, 177–196 (1988).
5. Conkright, M. E. *et al.* *World Ocean Atlas 2001: Objective Analyses, Data Statistics, and Figures, CD-ROM Documentation* (National Oceanographic Data Center, Silver Spring, 2002).
6. Warren, B. Why is no deep water formed in the North Pacific? *J. Mar. Res.* **41**, 327–347 (1983).
7. Haug, G. H., Sigman, D. M., Tiedemann, R., Pedersen, T. F. & Sarnthein, M. Onset of permanent stratification in the subarctic Pacific Ocean. *Nature* **400**, 779–782 (1999).
8. Hillenbrand, C.-D. & Fütterer, D. K. Neogene to Quaternary deposition of opal on the continental rise west of the Antarctic Peninsula, ODP Leg 178, Sites 1095, 1096 and 1101. *Proc. ODP Sci. Res.* **178**, 1–33 (2001).
9. Barker, P. F. & Kennett, J. P. Weddell Sea paleoceanography: Preliminary results of ODP Leg 113. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **67**, 75–102 (1988).
10. Altabet, M. A. & Francois, R. Sedimentary nitrogen isotopic ratio as a recorder for surface ocean nitrate utilization. *Glob. Biogeochem. Cycles* **8**, 103–116 (1994).
11. Andreev, A., Kusakabe, M., Honda, M., Murata, A. & Saito, C. Vertical fluxes of nutrients and carbon through the halocline in the western subarctic gyre calculated by mass balance. *Deep-sea Res. II* **49**, 5577–5593 (2002).
12. Narita, H. *et al.* Biogenic opal indicating less productive northwestern North Pacific during the glacial ages. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014320 (2002).
13. Sigman, D. M., Altabet, M. A., Francois, R., McCorkle, D. C. & Fischer, G. The $\delta^{15}\text{N}$ of nitrate in the Southern Ocean: Consumption of nitrate in surface waters. *Glob. Biogeochem. Cycles* **13**, 1149–1166 (1999).
14. Sigmon, D. E., Nelson, D. M. & Brzezinski, M. A. The Si cycle in the Pacific sector of the Southern Ocean: seasonal diatom production in the surface layer and export to the deep sea. *Deep-sea Res. II* **49**, 1747–1763 (2002).
15. Brzezinski, M. A. *et al.* A switch from $\text{Si}(\text{OH})_4$ to NO_3^- depletion in the glacial Southern Ocean. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL014349 (2002).
16. Lefevre, N. & Watson, A. J. Modeling the geochemical cycle of iron in the oceans and its impact on atmospheric CO_2 concentrations. *Glob. Biogeochem. Cycles* **13**, 727–736 (1999).
17. Francois, R. F. *et al.* Water column stratification in the Southern Ocean contributed to the lowering of glacial atmospheric CO_2 . *Nature* **389**, 929–935 (1997).
18. Gildor, H. & Tziperman, E. Physical mechanisms behind biogeochemical glacial-interglacial CO_2 variations. *Geophys. Res. Lett.* **28**, 2421–2424 (2001).
19. Boyle, E. A. & Keigwin, L. D. Deep circulation of the North Atlantic for the last 200,000 years: geochemical evidence. *Science* **218**, 784–787 (1982).
20. Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean. *Nature* **378**, 165–167 (1995).
21. Wang, Z. M., Mysak, L. A. & McManus, J. F. Response of the thermohaline circulation to cold climates. *Paleoceanography* **17**, doi:10.1029/2000PA000587 (2002).
22. Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. & Manabe, S. Response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393**, 245–249 (1998).
23. Toggweiler, J. R., Russell, J. Shifted westerlies caused low CO_2 during cold glacial periods. *Eos* **84** (Ocean Sci. Meet. Suppl.), OS21N-06 (2003).
24. Sarmiento, J. L. & Toggweiler, J. R. A new model for the role of the oceans in determining atmospheric $p\text{CO}_2$. *Nature* **308**, 621–624 (1984).
25. Froelich, P. N. *et al.* Biogenic opal and carbonate accumulation rates in the Subantarctic South Atlantic: The Late Neogene of Meteor Rise Site 704. *Proc. ODP Sci. Res.* **114**, 515–532 (1991).
26. Orsi, A. H., Whitworth, T. III & Nowlin, W. D. Jr On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res. I* **42**, 641–673 (1995).
27. Rea, D. K., Basov, I. A., Janacek, T. R. & Palmer-Julson, A. Introduction to Leg 145: North Pacific Transect. *Proc. ODP Init. Rep.* **145**, 5–7 (1993).
28. Burckle, L. H., Gersonde, R. & Abrams, N. Late Pliocene-Pleistocene paleoclimate in the Jane Basin region: ODP Site 697. *Proc. ODP Sci. Res.* **113**, 803–809 (1987).
29. Brewster, N. A. Cenozoic biogenic silica sedimentation in the Antarctic Ocean. *Geol. Soc. Am. Bull.* **91**, 337–347 (1980).
30. Diekmann, B., Fätker, M. & Kuhn, G. Environmental history of the south-eastern South Atlantic since the middle Miocene: Evidence from the sedimentological records of ODP Sites 1088 and 1092. *Sedimentology* **50**, 511–529 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the US NSF, Schweizer Nationalfonds, Deutsche Forschungsgemeinschaft, and by British Petroleum and Ford Motor Company through the Carbon Mitigation Initiative at Princeton University. This research used samples provided by the ODP. The ODP is sponsored by NSF and participating countries under the management of Joint Oceanographic Institutions. We thank T. F. Pedersen and K. Gordon for isotope analyses, M. Soon for analytical assistance, and C.-D. Hillenbrand for providing data from ODP Site 1096. J. D. Hays focused our attention on M. Winton's work on the climate implications of the nonlinear dependence of density on temperature. D. P. Schrag encouraged its pursuit as a general mechanism, and J. F. Adkins contributed the thoughts about the role of sea ice formation on the Antarctic shelf.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.S. (sigman@princeton.edu).

Hybrid fracture and the transition from extension fracture to shear fracture

Jonathan M. Ramsey* & Frederick M. Chester

Center for Tectonophysics, Department of Geology and Geophysics, Texas A&M University, College Station, Texas 77843-3115, USA

* Present address: Anadarko Petroleum Corporation, 1201 Lake Robbins Drive, The Woodlands, Texas 77380, USA

Fracture is a fundamental mechanism of material failure. Two basic types of brittle fractures are commonly observed in rock deformation experiments—extension (opening mode) fractures and shear fractures^{1,2}. For nearly half a century it has been hypothesized that extension and shear fractures represent end-members of a continuous spectrum of brittle fracture types^{3–6}. However, observations of transitional fractures that display both opening and shear modes (hybrids) in naturally deformed rock have often remained ambiguous, and a clear demonstration of hybrid fracture formation has not been provided by experiments⁴. Here we present the results of triaxial extension experiments on Carrara marble that show a continuous transition from extension fracture to shear fracture with an increase in compressive stress. Hybrid fractures form under mixed tensile and compressive stress states at acute angles to the maximum principal compressive stress. Fracture angles are greater than those observed for extension fractures and less than those observed for shear fractures. Fracture surfaces also display a progressive change from an extension to shear fracture morphology.

In the laboratory, a shear fracture is produced under a compressive stress state, shows displacement parallel to the failure surface, and forms at a 20°–40° angle to the maximum principal compressive stress direction, σ_1 (Fig. 1). An extension fracture forms under a tensile stress condition, displays an opening-mode displacement normal to the fracture surface, and propagates in an orientation parallel to the maximum and perpendicular to the minimum principal compressive stresses. Additional fracture types described in laboratory samples, such as wedging and longitudinal splitting during uniaxial compression, are often assumed to be forms of extension and shear fractures^{1,2,7,8}.

Within the compressive stress regime, the empirical Coulomb–Mohr failure criterion successfully predicts the brittle failure strength and orientation of shear fractures, at least to first order^{9,10}. As implied by Mohr's hypothesis, the orientation of the fracture surface is inclined to σ_1 at a fracture angle, θ , equal to $\pi/4 - \phi/2$ or $-\pi/4 + \phi/2$, where $\tan\phi$ is the slope of the failure envelope in Mohr space⁹. The failure envelope is commonly depicted as a continuous function that crosses the transition from a compressive to a tensile stress state, and thus includes uniaxial tensile failure^{11,12} (Fig. 1).

A theoretical basis for continuity across stress states is provided by the Griffith theory of fracture; the theory that forms the basis of modern linear elastic fracture mechanics^{1,2}. Griffith theory captures the essential physical processes of crack propagation, and provides an energy-based failure criterion for uniaxial tensile failure. From this theory Griffith derived a criterion for brittle failure under biaxial stress states, assuming a local tensile stress condition for crack propagation¹³. Further modifications treat failure under compression where crack closure and friction are important, and are generalized to treat failure under triaxial stress^{14–16}. The Griffith and modified-Griffith criteria predict a parabolic failure envelope for mixed tensile and compressive stress conditions that includes uniaxial tensile failure. The Griffith criteria have also been

combined with the Coulomb–Mohr criteria for compressive stress states to form a composite failure envelope^{17,18}. Such a continuous failure criterion is consistent with the hypothesis that a continuous transition from extension fracture to shear fracture exists and that hybrid fractures form under mixed tensile and compressive stress states (Fig. 1). Nevertheless, laboratory evidence documenting brittle failure in the form of a hybrid fracture produced under a mixed stress state is ambiguous^{8,19}.

The most comprehensive experimental study of brittle failure under mixed stress states is that of ref. 19. Twenty triaxial extension tests were conducted on five different rock types (granite, quartzite, diabase and two dolomites) using a dog-bone sample geometry. Dog-bone samples are cylinders of rock having a reduced central neck within which an axial tensile stress can be generated via axial extension under confining pressure, P_c (refs 19, 20) (Fig. 2). Although fifteen of the samples experienced an axial tensile stress at failure, the inconsistency in fracture orientation and failure strength precluded conclusions regarding the existence of a continuous transition from extension to shear fractures^{4,19}. The experimental data did, nonetheless, indicate that fracture under a mixed stress state may occur at a tensile stress that is less tensile than the uniaxial tensile strength.

To investigate hybrid fracture further we followed the general methodology of previous extension tests on dog-bone samples, but made key modifications to the sample geometry and jacket design to improve reproducibility. Samples were taken from the Lorano Bianco Carrara marble of Italy. The marble is a relatively homogeneous and isotropic, nearly pure calcite rock having a grain size of approximately 250 μm , extremely low porosity, and little crystal-lattice preferred orientation^{21,22}. Unlike the fillet-cut dog-bone geometry used previously^{19,20}, we used a large-radius notch-cut geometry produced by grinding along a single continuous circular arc (Fig. 2a). Photoelastic modelling and elasticity calculations indicate that the large-radius notch cut produces the most uniform stress condition, at least for uniaxial tensile loading²³.

Two experimental difficulties that may arise during extension of

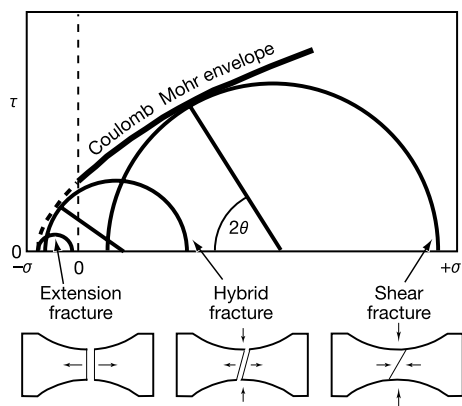


Figure 1 A representation using the Mohr diagram of the hypothesis that the brittle failure envelope and the transition from extension fracture to shear fracture is continuous. Below the Mohr diagram (plot of shear stress, τ , versus normal stress, σ) are schematics of the notch-cut dog-bone sample geometry displaying the three stress conditions and the associated types of fracture (extension, hybrid and shear). Three stress states satisfying the failure criterion are shown in the Mohr diagram by the point of tangency of the stress circles with the failure envelope. In the Mohr diagram, the fracture surface orientation is specified by the slope of the failure envelope at the point of tangency (see text). The predicted fracture angle, θ (measured between the fracture surface and the direction of the maximum principal compressive stress, σ_1), increases with increase in compressive normal stress. Hybrid fractures are thought to occur under mixed tensile and compressive stress conditions and form at small but non-zero θ .

dog-bone samples are the generation of bending forces and jacket effects. Our experiments were conducted in a gear-driven, liquid confining-media, triaxial apparatus that is very square and stiff ($9.47 \times 10^8 \text{ N m}^{-1}$), and capable of deforming samples 45 mm in diameter²⁴. Bending forces are minimized using this apparatus with precisely machined samples. Undesirable jacket effects were obviated by using layered jackets composed of an inner layer of latex or copper and an outer jacket of polyolefin separated by plasticene modelling clay²². The outer polyolefin is impermeable to the confining media and the clay conveys confining stress to the necked portion of the sample. The clay is necessary because the polyolefin supports relatively large elastic stresses and induces undesirable loading of the sample if it is in direct contact with the rock. An inner layer of latex restricts the clay from intruding into rock pores before fracture, and allows very accurate measurement of failure strength because the latex is very weak. Detailed observation of the fracture surfaces is precluded by use of latex because the latex ruptures after failure and the clay intrudes into the sample. Accordingly, a parallel suite of experiments were conducted using jackets with an inner layer of copper. The copper was sufficiently strong to ensure preservation of the fracture surfaces.

Thirty-one extension experiments were completed at room temperature (22 °C) and an axial extension rate of 2×10^{-2}

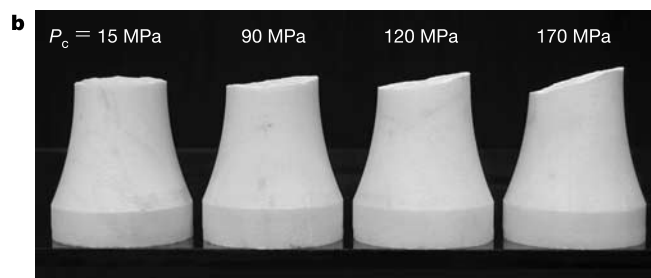
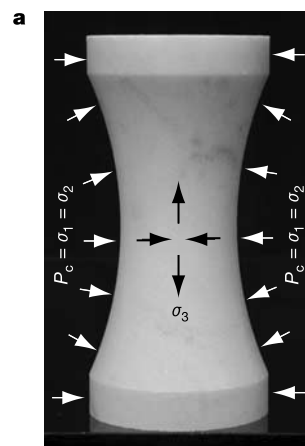


Figure 2 Photographs of the undeformed and deformed notch-cut dog-bone samples of Carrara marble. **a**, Samples are cylindrical and the notch cut is made with a 91.5-mm radius grinding wheel. The final sample is 90 mm long, with 43 mm diameter at the ends, and 28 mm diameter in the central portion of the notch. Arrows indicate the confining pressure, P_c , applied along boundaries of the cylindrical sample that generate differential stress as the sample ends are unloaded. In the central portion of the notch cut, the minimum principal compressive stress, σ_3 , is in the direction of the cylinder axis and may be tensile (negative). The maximum and intermediate principal stresses are the same magnitude (equal to P_c) and are oriented perpendicular to the cylinder axis. **b**, Representative samples showing the progressive change in fracture orientation with increase in P_c . Fractures form in the samples at the neck (central portion of the notch); the other (matching) halves of the samples are not included in the photograph.

mm s⁻¹. The results of 18 experiments with copper jackets and 13 experiments with latex jackets were combined to define failure strength, fracture orientation, and fracture surface morphology as a function of confining pressure (Table 1). Fracture surfaces are approximately planar at the sample scale, but are complex at the grain scale. Fracture orientation is specified by the maximum angle between the fracture surface and the plane normal to the cylinder axis (which contains the σ_1 direction, Fig. 2a). Overall, there is an increase in fracture angle with confining pressure (Figs 2b and 3a). The axial stress in the central region of the neck, which is the minimum principal compressive stress, σ_3 , is determined from measurements of axial force and confining pressure^{19,20}. A tensile axial stress (negative σ_3) was achieved at failure in all tests run at confining pressures less than 130 MPa (Fig. 3b).

On the basis of fracture orientation, fracture surface morphology, and stress state at failure, three major classes of fractures are identified within the total population. Fractures formed at the lowest confining pressures (7.5 and 60 MPa) are oriented approximately perpendicular to σ_3 (parallel to σ_1 ; Fig. 3a), form when σ_3 equals -7.8 MPa (Fig. 3b; close to the uniaxial tensile strength of 6.9 MPa determined²⁵ for Carrara marble), display a surface morphology consistent with an opening-mode displacement, and may be classified as extension fractures^{1,2}. The fracture surfaces are somewhat undulating at the sample scale, and display discrete, highly reflective intragranular cleavages. In most cases the cleavage surfaces in neighbouring grains are not coplanar, so the surfaces are rough at the grain scale.

Fractures formed at confining pressures of 130 to 170 MPa are inclined 13° to 22° from σ_1 (Fig. 3a), form under a compressive stress state (Fig. 3b), display surface features consistent with shear displacement, and may be classified as shear fractures^{1,2}. The surfaces are covered with a powder of comminuted grains and contain short grooves. The grooves are parallel to the maximum inclination (that is, in the direction of the maximum shear traction) and record the direction of slip. Cleaved crystals, as seen on extension fracture surfaces, are not present. In cross-sectional view the marble is discoloured adjacent to the fracture surface, which probably reflects a microfractured zone associated with the through-going macroscopic fracture.

Fractures formed at confining pressures between 70 to 120 MPa are inclined 3° to 13° from σ_1 (Fig. 3a), form under a mixed tensile and compressive stress state (Fig. 3b), and display transitional surface characteristics expected for hybrid fractures^{3,4}. Specifically, the surfaces exhibit patches of discrete, reflective, cleaved crystals between areas of comminuted material with slip lineations. The morphology of the fracture surfaces show a strong correlation to confining pressure. Fracture surfaces formed at the lowest pressure

display the highest total area of cleavage planes relative to area of comminuted material. Conversely, fracture surfaces formed at the highest pressures display the lowest total area of reflective cleavage planes relative to area of comminuted material. The thickness of the discoloured zone adjacent to the fracture surface also increases with increasing confining pressure.

Our observations of the progressive increase in fracture angle with confining pressure, from extension fracture at low pressure to shear fracture at high pressure, provides the first convincing laboratory evidence for the existence of hybrid fractures that constitute a continuous transition from extension to shear fractures. The results reported herein support previous interpretations that some natural fractures may be hybrids or conjugate shear fractures with narrow dihedral angles. The surface morphology of the hybrid fracture is compatible with the step-tread geometry of the fractures

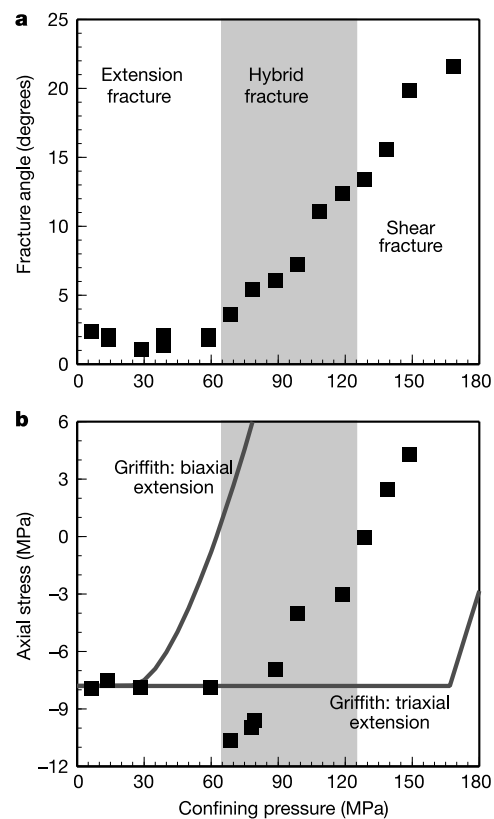


Figure 3 Fracture angle and fracture strength as a function of confining pressure. Extension, hybrid and shear fracture are distinguished on the basis of fracture orientation and surface morphology (see text). **a**, Fracture angle increases monotonically with confining pressure for hybrid and shear fractures; fracture angle is constant for extension fractures. The fracture angle is the maximum angle between the fracture surface and the plane normal to the cylinder axis; the plane normal to the cylinder axis contains σ_1 and σ_2 directions. Orientation of the fracture surface was determined by the best-fitting plane to 24 evenly distributed measurements of surface height (along the cylinder axis); measurements are made using a drag profilometer with accuracy of 50 μ m. **b**, Fracture strength increases approximately linearly with confining pressure for hybrid and shear fracture, and is approximately constant for extension fracture. Fracture strength is represented by the magnitude of the minimum principal stress in the neck of the sample at the ultimate load-bearing capacity. There is an abrupt (possibly discontinuous) change in fracture strength at the transition from extension fracture to hybrid fracture at a P_c between 60 and 70 MPa. Fracture strength predicted by the Griffith criteria for biaxial ($\sigma_1 > \sigma_3$, $\sigma_2 = 0$) and triaxial-extension ($P_c = \sigma_1 = \sigma_2 > \sigma_3$) states-of-stress are shown by solid lines. The Griffith criterion produces a parabolic failure envelope in Mohr space.

Table 1 Combined results of Carrara marble triaxial extension experiments

$P_c = \sigma_1$ (MPa)	$\sigma_1 - \sigma_3$ (MPa)	σ_3 (MPa)*	Fracture angle (degrees)†	Fracture morphology
7.5	15.4	-7.9	2.4	Extension
15	22.5	-7.5	1.8, 2.1	Extension
30	37.8	-7.8	1.1	Extension
40	—	—	1.4, 2.1	Extension
60	67.9	-7.8	1.9, 2.1	Extension
70	80.6	-10.6	3.7	Hybrid
80	89.7, 89.8	-9.7, -9.8	5.5	Hybrid
90	96.9	-6.9	6.1	Hybrid
100	104.0	-4.0	7.3	Hybrid
110	—	—	11.1	Hybrid
120	123.0	-3.0	12.4	Hybrid
130	130.0	0	13.4	Shear
140	137.5	2.5	15.6	Shear
150	145.7	4.3	19.9	Shear
170	—	—	21.6	Shear

* Determined from latex jacket experiments.

† Determined from copper jacket experiments.

classified as hybrids in the Mount Desert Island granite^{4,26}. This geometry is consistent with fracture formation through linkage of an array of en echelon, interacting, extension cracks²⁷.

Although a continuous transition from extension to shear fracture was hypothesized on the basis of Griffith theory and the extrapolation of the Mohr envelope into the tensile field, our data do not fully support such an approach. Contrary to Coulomb–Mohr analysis, the fracture angles observed for hybrid and shear fractures are systematically less than predicted by the slope of the failure envelope. In addition, the failure strengths of the marble for mixed tensile and compressive stresses are not consistent with a continuous, parabolic failure envelope such as that predicted by the Griffith criterion (Fig. 3b). Additional study is needed to determine the specific processes that govern formation of hybrid fractures, but given that hybrid fractures are transitional to extension and shear fractures, it is likely that a complete understanding will require an appropriate fracture mechanics theory that considers the mechanical interaction of stepped cracks^{4,27}. □

Received 1 December 2003; accepted 13 January 2004; doi:10.1038/nature02333.

1. Jaeger, J. C. & Cook, N. G. W. *Fundamentals of Rock Mechanics* (Chapman and Hall, London, 1979).
2. Paterson, M. S. *Experimental Rock Deformation—The Brittle Field* (Springer, Berlin, 1978).
3. Hancock, P. L. in *Geology in the Real World—the Kingsley Dunham Volume* (eds Nichol, I. & Nesbitt, R. W.) 155–164 (Institution of Mining and Metallurgy, London, 1986).
4. Engelder, T. Transitional-tensile fracture propagation: a status report. *J. Struct. Geol.* **21**, 1049–1055 (1999).
5. Muehlberger, W. R. Conjugate joint sets of small dihedral angle. *J. Geol.* **69**, 211–219 (1961).
6. Sibson, R. H. Brittle-failure controls on maximum sustainable overpressure in different tectonic regimes. *Am. Assoc. Petrol. Geol. Bull.* **87**, 901–908 (2003).
7. Griggs, D. T. & Handin, J. W. Observations on fracture and a hypothesis of earthquakes. *Geol. Soc. Am. Mem.* **79**, 347–364 (1960).
8. Hoskins, E. R. The failure of thick-walled hollow cylinders of isotropic rock. *Int. J. Rock Mech. Min. Sci.* **6**, 99–125 (1969).
9. Handin, J. On the Coulomb–Mohr failure criterion. *J. Geophys. Res.* **74**, 5343–5348 (1969).
10. Mogi, K. Effect of the intermediate principal stress on rock failure. *J. Geophys. Res.* **72**, 5117–5131 (1967).
11. Leon, A. Über die Rolle des Trennbruches im Rahmen der Mohrschen Anstrengungshypothese. *Bauingenieur* **15**, 318–321 (1934).
12. Nadai, A. *Theory of Flow and Fracture of Solids* (McGraw Hill, New York, 1934).
13. Griffith, A. A. Theory of rupture. *Proc. Int. Congr. Appl. Mech.* **1**, 55–63 (1924).
14. McClintock, F. A. & Walsh, J. B. in *Proc. 4th US Nat. Congr. Appl. Mech.* Vol. II, 1015–1021 (Am. Soc. Mech. Eng., New York, 1962).
15. Murrell, S. A. F. The effect of triaxial stress systems on the strength of rocks at atmospheric temperatures. *Geophys. J. R. Astron. Soc.* **10**, 231–281 (1965).
16. Murrell, S. A. F. & Digby, P. J. The theory of brittle fracture initiation under triaxial stress conditions. I and II. *Int. Geophys. J. R. Astron. Soc.* **19**, 309–334 and 499–512 (1970).
17. Brace, W. F. An extension of the Griffith theory of fracture to rocks. *J. Geophys. Res.* **65**, 3477–3480 (1960).
18. Secor, D. T. Role of fluid pressure in jointing. *Am. J. Sci.* **263**, 633–646 (1965).
19. Brace, W. F. in *State of Stress in the Earth's Crust* (ed. Judd, W. R.) 111–180 (American Elsevier, New York, 1964).
20. Schock, R. N. & Louis, H. Strain behavior of a granite and a greywacke sandstone in tension. *J. Geophys. Res.* **87**, 7817–7823 (1982).
21. Pieri, M., Burlini, L., Kunze, K., Stretton, I. & Olgaard, D. Rheological and microstructural evolution of Carrara marble with high shear strain: results from high temperature torsion experiments. *J. Struct. Geol.* **23**, 1393–1413 (2001).
22. Ramsey, J. M. *Experimental Study of the Transition from Brittle Shear Fractures to Joints* (Texas A&M Univ. Press, College Station, TX, 2003).
23. Neuber, H. *Theory of Notch Stresses* (McGraw-Hill, New York, 1937).
24. Handin, J., Friedman, M., Logan, J. M., Pattison, L. J. & Swolfs, H. S. in *Flow and Fracture of Rocks* (eds Heard, H. C., Borg, I. Y., Carter, N. L. & Raleigh, C. B.) 1–28 (AGU Geophys. Monogr. 16, American Geophysical Union, 1972).
25. Jaeger, J. C. & Hoskins, E. R. Stresses and failure in rings of rock loaded in diametral tension or compression. *Br. J. Appl. Phys.* **17**, 685–692 (1966).
26. Engelder, T. The analysis of pinate joints in the Mount Desert Island granite: implications for post-intrusion kinematics in the coastal volcanic belt, Maine. *Geology* **17**, 564–567 (1989).
27. Reches, Z. & Lockner, D. Nucleation and growth of faults in brittle rocks. *J. Geophys. Res.* **99**, 18159–18174 (1994).

Acknowledgements We thank M. Finn, A. K. Kronenberg and J. S. Chester for discussions and suggestions, J. N. Magouirk and E. C. Powell for assistance in the laboratory, and J. S. Chester for reviews of the manuscript. This work was partially supported by US NSF.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.M.C. (chesterf@geo.tamu.edu).

Coral mucus functions as an energy carrier and particle trap in the reef ecosystem

Christian Wild¹, Markus Huettel², Anke Klüeter³, Stephan G. Kremb⁴, Mohammed Y. M. Rasheed⁵ & Bo B. Jørgensen¹

¹Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

²Department of Oceanography, Florida State University, Tallahassee, Florida 32306-4320, USA

³Australian Institute of Marine Science, PMB No. 3, Townsville MC, Queensland 4810, Australia

⁴Alfred Wegener Institute for Polar and Marine Research, Department Benthic Ecosystems, Am Handelshafen 12, D-27570 Bremerhaven, Germany

⁵Marine Science Station, University of Jordan and Yarmouk University, PO 195, Aqaba, Jordan

Zooxanthellae, endosymbiotic algae of reef-building corals, substantially contribute to the high gross primary production of coral reefs¹, but corals exude up to half of the carbon assimilated by their zooxanthellae as mucus^{2,3}. Here we show that released coral mucus efficiently traps organic matter from the water column and rapidly carries energy and nutrients to the reef lagoon sediment, which acts as a biocatalytic mineralizing filter. In the Great Barrier Reef, the dominant genus of hard corals, *Acropora*, exudes up to 4.8 litres of mucus per square metre of reef area per day. Between 56% and 80% of this mucus dissolves in the reef water, which is filtered through the lagoon sands. Here, coral mucus is degraded at a turnover rate of at least 7% per hour. Detached undissolved mucus traps suspended particles, increasing its initial organic carbon and nitrogen content by three orders of magnitude within 2 h. Tidal currents concentrate these mucus aggregates into the lagoon, where they rapidly settle. Coral mucus provides light energy harvested by the zooxanthellae and trapped particles to the heterotrophic reef community, thereby establishing a recycling loop that supports benthic life, while reducing loss of energy and nutrients from the reef ecosystem.

Zooxanthellae provide a substantial part of the energy requirement of their hosts by transferring photosynthetically fixed carbon to the coral⁴. High arabinose contents in the carbohydrate fraction of coral mucus indicate that much of the fixed carbon is released as mucus, because arabinose is not usually a constituent of animal cells⁵. Many hard and soft corals release mucus continuously, with a species-specific composition⁶. Known functions of this release are protection against fouling⁶, desiccation during air exposure at extreme low tide⁷ and sedimentation⁸.

The adhesive mucus has the ability to trap particulate matter from the water and the light-collecting coral surfaces. Mucus and trapped particles are transported over the coral surface by ciliary currents and are released to the water⁹ at a magnitude such that mucus can dominate suspended matter around reefs^{10,11}. Our measured carbon to nitrogen ratios of 5–14 point towards the potential value of coral mucus as a food source; however, its role in the cycling of matter in coral reef environments is largely unknown.

We investigated release rates, chemical composition, contents and fate of mucus produced by *Acropora*, the coral genus with the highest area of coverage (15.8%) on the reef rim (crest plus slope) of Heron Island (23° 27' S, 151° 55' E), a ring-shaped platform reef in the Great Barrier Reef, Australia. Submerged *Acropora* released 1.7 l of mucus per m² of reef area per day, excluding the rapidly dissolving mucus fraction. *In situ* measurements carried out after air exposure, a regular event caused by extreme low tides¹² and

occurring on about 6 d month^{-1} for 2 h d^{-1} at Heron Island, indicated that reef rim *Acropora* can release 4.81 of mucus per m^2 of reef area per day. This corresponds to $10\text{--}21 \text{ mmol}$ of particulate organic carbon (POC), $1.5\text{--}1.8 \text{ mmol}$ of nitrogen and $0.08\text{--}0.18 \text{ mmol}$ of phosphorus.

More than half ($56\text{--}80\%$) of the released coral mucus immediately dissolves in the sea water, and this dissolved fraction provides a food source for planktonic bacteria^{6,13}. The release of coral dissolved organic carbon stimulates microbial metabolism and growth¹⁴. Previous studies have agreed that coral mucus, primarily its energy- and nutrient-rich components such as proteins, triglycerides and wax ester, can be partly degraded in the water column^{13,15}. Our mucus respiration experiments, conducted within 60 min of mucus release, showed high consumption rates of O_2 of $130\text{--}445 \mu\text{mol l}^{-1} \text{ d}^{-1}$, as compared with $5\text{--}41 \mu\text{mol l}^{-1} \text{ d}^{-1}$ in the surrounding sea water. This is caused by 100-fold higher bacterial numbers in the mucus (2.7×10^7 versus $3.0 \times 10^5 \text{ cells ml}^{-1}$, $n = 3\text{--}6$).

The less-soluble fraction ($20\text{--}44\%$) of the exuded gel-like mucus forms transparent filaments and strings (Fig. 1a) that detach from the coral branches and subsequently combine to mucus floes. Positive buoyancy, caused by enclosed gas bubbles and lipids, results in a slow ascent of these floes. Passing through the water column, their sticky surface traps bacteria, algal cells and small carbonate particles (Fig. 1b). Enriched mucus aggregates then accumulate at the water surface and form opaque films that are concentrated by currents and winds to produce whitish mucus threads. The accumulation and subsequent fusion of these threads generate mucus floats

with a thickness of $1.5\text{--}3.0 \text{ cm}$ (Fig. 1c) and a width and length of up to 2 and 10 m , respectively. These cohesive floats trap larger particles, including filamentous algae, small zooplankton and sand grains (Fig. 1d). Two hours after low tides with air exposure of the corals, the gel-like transparent mucus released from the corals has turned into a yellow, bubbly mass that smells strongly of dimethyl sulphide and contains on average $2,250 \pm 230 \text{ mM POC}$, $220 \pm 40 \text{ mM nitrogen}$ and $2 \pm 0.1 \text{ mM phosphorus}$ (carbonate grains excluded, $n = 4\text{--}8$).

Roller table experiments¹⁶ using sea water or coral mucus each mixed with a suspension containing zooxanthellae, fine carbonate grains ($<10 \mu\text{m}$) or bacteria showed that, after $3\text{--}14 \text{ h}$ of slow rotation, aggregates had formed only in the cylinders containing mucus. This shows that mucus functions as a trap for particles and causes the formation of aggregates. Because of compaction, the resulting mucus-particle aggregates had a volume that was 300-fold smaller than the volume of the mucus added initially. Likewise, natural mucus floats loaded with particles trapped from the reef waters (Fig. 1c) showed a reduced average water content of only $60 \pm 1 \text{ vol.}\%$ ($n = 6$). The enrichment factors presented in Table 1, with correction for compaction, lead to an eightfold and ninefold increase in carbon and nitrogen, respectively, but to no increase in phosphorus relative to the original exuded mucus.

Continuous compaction and accumulation of resuspended carbonate grains (Fig. 1d) gradually decrease the positive buoyancy of the mucus floats. Water movement of the incoming tide transports the mucus into the lagoon and disintegrates the floats. Subsequent release of trapped air bubbles initiates a rapid sinking ($4\text{--}8 \text{ cm s}^{-1}$)

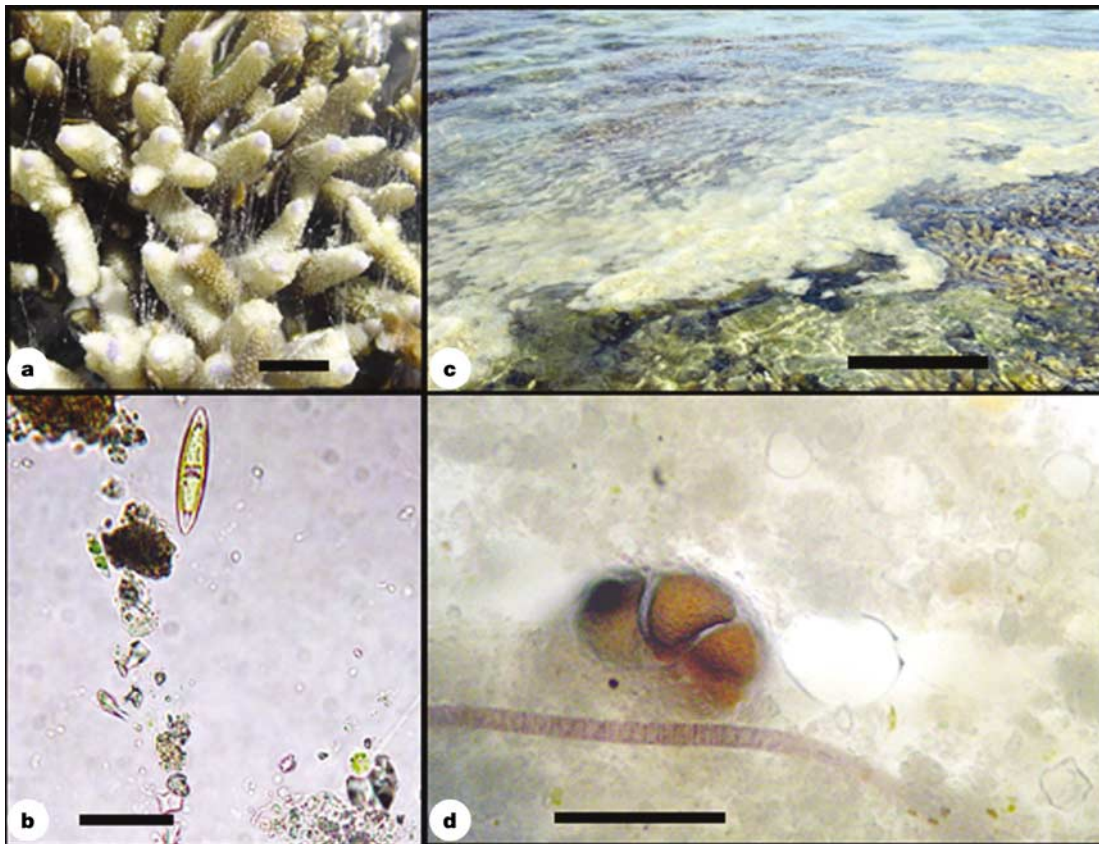


Figure 1 Changes in coral mucus during its ageing process. **a**, Image shows mucus strings, freshly produced by *Acropora* sp., attached to the coral. **c**, Less than 2 h later, large mucus floats drift on the water surface in the direction of the lagoon. **b**, **d**, Microscopic observations indicate that there are considerable differences in content between freshly produced mucus and mucus floats. In fresh mucus (**b**), a diatom (top centre), a dinoflagellate (bottom right) and a few carbonate grains are seen attached to the

mucus strings, whereas aged mucus (**d**) shows an accumulation of a large foraminifer (brown, centre) and a filamentous alga (bottom), in addition to heavy loads of carbonate grains. See Table 2 for a comparison of the compositional analyses of both mucus stages, and the enrichment factors for the measured parameters. Scale bars, 1 cm (**a**), 1 m (**c**), $50 \mu\text{m}$ (**b**, **d**).

to the sandy bottom of the reef lagoon. A series of plankton nets (diameter, 26 cm; mesh, 500 μm ; each net fixed 20 cm above the sediment), placed along the transect shown in Fig. 2, caught mucus aggregates close to the lagoon floor, revealing sedimentation of these aggregates more than 150-m inward of the reef crest (Fig. 2b).

The lagoon sediment is characterized by medium- to coarse-grained carbonate sands with high permeability ($1.2 \pm 0.4 \times 10^{-10} \text{ m}^2$, $n = 7$), high porosity ($44 \pm 2 \text{ vol.}\%$, $n = 8$) and a large specific surface area of the sand grains ($0.18\text{--}0.74 \text{ m}^2 \text{ g}^{-1}$). Although the settling mucus aggregates are a desirable food source for bottom dwellers⁷, the dissolved fraction of coral mucus is carried into the permeable sands by interfacial water flows caused by wave pumping¹⁷ and tide-induced differences in water level (up to 0.5 m at Heron Island; data not shown) between the reef lagoon and the deeper water body outside the reef^{18,19}. Our tracer measurements showed that at least $264 \text{ l m}^{-2} \text{ d}^{-1}$ are pumped through the upper layers of the lagoon sands, excluding the pumping owing to tidal changes in water levels. At an average depth of the Heron Island lagoon of 1.7 m, this implies that more than 15.5% of the lagoon water is filtered through the sediment each day. These sedimentary and hydrodynamic properties, in combination with a dense population of benthic bacteria (5.8×10^8 to $6.3 \times 10^9 \text{ cells cm}^{-3}$), convert the reef lagoon sands into a large filter system that removes coral mucus from the water column and degrades this organic matter in the upper sediment layers.

Flux measurements with 4–6 stirred benthic chambers, placed simultaneously on the permeable lagoon sands, show a rapid decomposition of freshly produced coral mucus by sedimentary bacteria²⁰. Experimental addition of mucus to the chamber water enhanced consumption of O_2 immediately by 16–41% (Table 1). Oxygen consumption owing to mucus degradation in the water column accounted only for less than 10% of the total measured consumption of O_2 , showing that the benthic community decomposed more than 90% of the added mucus. The increased benthic O_2 uptake of $17\text{--}51 \text{ mmol m}^{-2} \text{ d}^{-1}$ suggests that at least 7% of the carbon added as coral mucus was turned over per hour²⁰. Using our measured carbon/nitrogen/phosphorus ratio of 72/6/1 for coral mucus, and assuming that all nutrients contained in older mucus and associated particles reach the lagoon sediment to be degraded, we arrive at potential release rates of $1.92 \text{ mmol m}^{-2} \text{ d}^{-1}$ for

nitrogen and $0.04 \text{ mmol m}^{-2} \text{ d}^{-1}$ for phosphorus (Fig. 3).

To calculate the transport rates of energy and nutrients as coral mucus to the lagoon sediments, we used a satellite picture of Heron Island (Fig. 2a) and derived a total coral covered surface area of 6.9 km^2 and a lagoon sediment surface area of 19.5 km^2 . The genus *Acropora* on the reef rim exuded at least 4.9 t of dry mass mucus per day (a conservative estimate for a day without air exposure) corresponding to 1.8 t of carbohydrate polymers, 1.0 t of protein and 0.2 t of lipids (using the *Acropora* carbohydrate/protein/lipid ratio from ref. 5). Owing to the circular shape of the reef, roughly 50% of the dissolved and undissolved mucus drifts into the lagoon. After aerial exposure of the reef rim, the incoming tide may push up to 100% of the mucus into the lagoon, owing to the pressure gradient between the low water level of the lagoon and the higher level outside the reef. In the lagoon, energy and nutrients contained in the mucus and trapped particles are transferred to the water column²¹ and to benthic food chains feeding bacteria, benthic invertebrates and fish (refs 7, 11, 22, 23, and Fig. 3).

The relative importance of coral mucus as a carrier of energy and

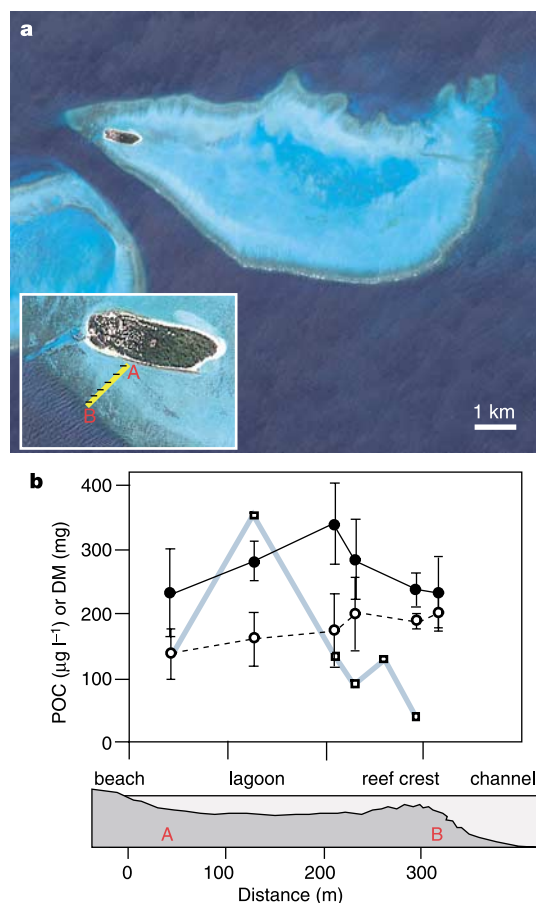


Figure 2 Transect studies at Heron Island. **a**, Aerial view of the Heron Island reef. Reproduced with permission from Space Image. The brownish-grey reef rim encloses the light-coloured sandy lagoon. Yellow line in inset shows the position of the sampling transect A–B, with black bars indicating the six sampling stations. **b**, POC is highest at the lagoon-ward edge of the reef crest during flooding after coral exposure (filled circles), and is almost double the concentration at ebb tide (open circles). Concentrations of POC were found to be significantly higher in high-tide than in low-tide conditions ($P < 0.001$), with no significant interaction between tide condition and shore distance ($P > 0.2$; two-way analysis of variance). Plankton nets caught the main load of mucus aggregates in the lagoon (grey line and open squares). Microscopic examination of the aggregates caught in the nets verified that these were indeed mucus aggregates similar to those shown in Fig. 1c. DM, dry mass.

Table 1 Benthic degradation of coral mucus studied with stirred benthic chambers

Date of <i>in situ</i> experiment	Control ($\text{mmol O}_2 \text{ m}^{-2} \text{ d}^{-1}$) ^a	Mucus addition ($\text{mmol O}_2 \text{ m}^{-2} \text{ d}^{-1}$) ^a	Increase (%)
4 February 2001	52, 58, 80 (15)	99, 111, 133 (17)	41
15 January 2002	64, 67	71, 100	23
21 January 2002	82, 92	95, 113	16
25 January 2002	49, 92	106, 110	35

^aData are the O_2 consumption, given as mmol of O_2 per m^2 per day (s.d. in parentheses), calculated from the decrease in O_2 concentration in the enclosed water of each chamber (volume 6.8–7.0 l). The addition of 140–300 ml of freshly collected gel-like coral mucus to the chamber water (corresponding to 500–1,500 μg of POC per litre, compared with 171–488 μg of POC per litre measured in waters from the Heron Island reef flat, $n = 56$) caused a significant increase in O_2 consumption in the dark benthic chambers (two-sided *U*-test after Wilcoxon, Mann and Whitney, $\alpha = 0.002$) relative to the controls. Less than 10% of this increase could be attributed to the water column degradation of mucus, indicating that >90% of the added mucus was degraded in the permeable reef sands.

Table 2 Composition of freshly exuded coral mucus and aged mucus floats

	Freshly exuded mucus ^a	Mucus floats ^a	Enrichment factor	<i>n</i>
Dry mass (g l^{-1})	0.3 ± 0.1	107 ± 45	357	6
POC (mmol l^{-1})	2.3 ± 1.2	$2,253 \pm 227$	980	8
PN (mmol l^{-1})	0.2 ± 0.1	218 ± 36	1,089	8
C/N ratio	12 ± 3	11 ± 1	—	8
Total P ($\mu\text{mol l}^{-1}$)	32 ± 13	$1,835 \pm 71$	58	4
Chl <i>a</i> ($\mu\text{g l}^{-1}$)	3.7 ± 0.5	$8,134 \pm 2,809$	2,198	8

^aData are the mean $\pm$ standard deviation.

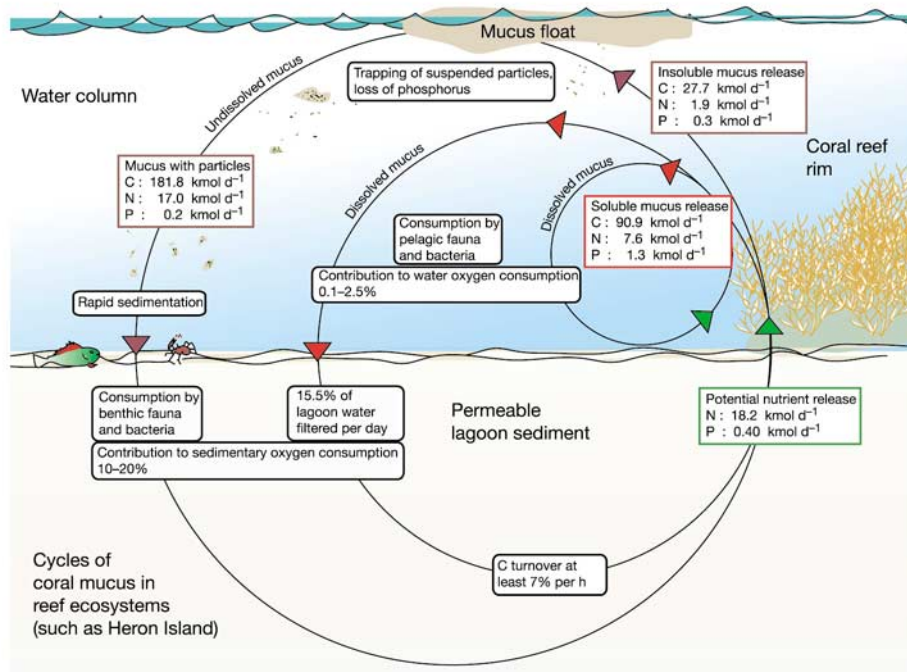


Figure 3 Proposed cycle of coral mucus for the Heron Island reef ecosystem. Mucus is produced in considerable quantities by branching hard corals covering a large proportion (29% coverage, as compared with 40% coverage by all hard corals) of the reef rim area (6.9 km²). Tidal currents dissolve 56–80% of the mucus, detach the mucus strings from the corals and transport at least 50% of dissolved and undissolved mucus into the lagoon (total area 19.5 km²). Here, the dissolved fraction fuels bacteria and algae of the

planktonic food chain and is also transported into the permeable reef sands, enhancing benthic metabolism. The effect of tidal pumping is not included in the water filtration calculations. Floating mucus aggregates trap suspended particles, sink, and fuel the benthic food chain. Pelagic and benthic mineralization of mucus as well as trapped particles returns regenerated nutrients to the corals and autotrophic reef organisms.

nutrients to consumers in the reef lagoon increases with the ratio between the coral-covered area and the lagoon area. The following estimates are based only on the undissolved mucus fraction; trapped particles and increased rates of mucus production as a consequence of aerial exposure were not included in these calculations. Our estimate suggests that, in the Heron Island reef, *Acropora* mucus production contributes at least 0.9–1.7% to the measured sedimentary decomposition of organic carbon. Presuming a comparable production of mucus by the other branching hard corals, which account for 73% of all hard corals growing on the Heron Island reef rim, this contribution increases to 1.3–2.4%. Including the trapped particulate matter, mucus-particle aggregates provide 10–20% of the total organic carbon that is metabolized by the sedimentary community, emphasizing the importance of this process for the energy and nutrient supply of the lagoon sands. On account of the trapping mechanism, nitrogen can be returned to the corals that was lost by mucus release (Fig. 3). In addition, without the presence of the 'mucus trap' in the reef water, many suspended particles would pass through the reef without being available for the benthic food chain²⁴.

In reefs with enclosed lagoons, coral reef and lagoon sediments together form a recycling entity linked by coral mucus as an important carrier of energy and nutrients (Fig. 3). By producing and shielding an enclosed or landward lagoon, coral reefs generate a biocatalytic filter system that accumulates and retains energy and nutrients in the reef ecosystem. A coral barrier protecting the lagoon provides the hydrodynamic conditions and tide-induced pressure gradients, which are necessary to increase the volume of water and particles filtered by the lagoon sands and the reef framework. Benthic organisms living on and in the lagoon sands and framework²⁵ can degrade organic matter from the water forced through these permeable structures and return nutrients to the water that are essential for autotrophic growth.

This tight recycling mechanism and the trapping function of mucus thus counteract the release of almost 50% of the net fixed carbon by the corals^{2,3} and may contribute to explain the high gross primary production that characterizes coral reefs growing in oligotrophic oceans. □

Methods

Mucus collection

Coral mucus was collected from the staghorn corals *Acropora millepora*, *Acropora pulchra*, *Acropora nobilis* and *Acropora aspera*. Individual coral colonies (maximum diameter, 30 cm), anchored with their lower branches in the carbonate sands, were lifted, exposed to air and subsequently released 0.1–0.3 l of liquid mucus. We discarded the initial 30 s of mucus release containing water and afterwards collected the dripping mucus into a container for 2 min.

Carbon and nitrogen analyses

Mucus aliquots (5–10 ml) were filtered onto pre-combusted GF/F filters (Whatman) and frozen at –20° C. Carbonate grains were removed by exposing the filters to fuming HCl. We then measured POC and particulate nitrogen (PN) using an NA1500 element analyser (Fisons).

Mucus solubility

One of two identical volumes of freshly collected gel-like mucus from *Acropora* was diluted (1:1 with 0.2-µm filtered sea water), and then each of the two volumes was added to a 5-ml graded cylinder with saturated NaCl solution containing rhodamine dye and mixed ($n = 11$ replicates for each treatment). Undissolved mucus rose to the surface, where it could be detected and quantified owing to its unstained appearance. The percentage mucus solubility was calculated as the ratio of the volume of the dissolved mucus fraction to the total volume of the gel-like mucus.

Quantification of mucus release

Mucus release of *Acropora* during naturally occurring exposure to air at low tide was quantified in five *in situ* experiments. For each experiment, individual colonies were transferred to three 10-l containers without exposure to air; three containers with only sea water were used as controls. During ebb tide, a gauze-covered (125-µm) opening permitted the water level in the containers to fall at the same rate as in the reef, thus finally exposing the incubated colonies to air. Exposure times were the same as those of the corals on the reef rim (15–105 min). The flooding tide slowly re-filled the containers through the

gauze-covered openings. As soon as the corals were submerged again, the openings were closed and, for exactly 3 min, a water current equivalent to ambient currents was generated by an electric pump, causing the detachment of mucus from the corals. We then removed the corals and thoroughly mixed the water before collecting samples (50–150 ml, $n = 3$ for each analysis) from all six containers for dry mass, POC and PN analysis.

Mucus production of fully submerged colonies of *A. millepora* and *A. aspera* was quantified as described²⁶. In brief, the small *Acropora* colonies ($n = 8$ in each of two independent experiments) were carefully transferred into sea water-filled glass beakers (500–2,000 ml) and incubated separately for 4–6 h under *in situ* conditions. Three additional beakers without coral and with unfiltered sea water were incubated as controls. At the end of the incubation time, the corals were removed from the glass beakers and the remaining water was mixed. Water samples from each of the eight beakers were taken for dry mass, carbon and nitrogen determination as described above. The results of all experiments were related to the three-dimensional surface area of the coral branches, which was assessed by a wax-coating method²⁷.

Daily mucus release of *Acropora* corals per m² of reef was calculated by using a factor of 3.8 to recalculate the three-dimensional surface area of the corals to the two-dimensional reef rim area covered by *Acropora*. This recalculation factor was derived from the ratio of measured coral surface area to the two-dimensional area (length $\times$ width) of the same coral colonies ($n = 15$). We calculated the coral-covered area on the Heron Island Reef rim from a satellite picture (Fig. 2a).

Hard coral coverage

Spatial distributions of hard corals on the reef rim were determined along nine randomly placed line transects. A counting grid (1 m²) was moved ten times along each line transect (10 m in length). Hard corals were identified down to species level, permitting calculation of the percentage cover of *Acropora*.

Oxygen consumption in suspended mucus

Freshly collected *Acropora* mucus and separately ambient sea water were incubated in 30-ml bottles in the dark and at *in situ* temperature (26–29 °C). Incubations started within 60 min of mucus collection or, if this was not possible, coral mucus was kept at 4 °C until incubation. We measured the O₂ concentrations in the incubated solutions simultaneously in time series by the Winkler titration method, and derived O₂ consumption rates from linear regression with at least four data points. Measurements were repeated on five different mucus samplings.

Water level measurements

Tide-induced water level differences between the lagoon and the water surrounding the Heron Reef were measured with a hydraulic potentiometer during two spring low tides.

In situ benthic chamber incubations

We studied benthic degradation of coral mucus in four independent *in situ* incubations using stirred advection chambers similar to those described in ref. 28. Details of these chamber experiments are found in ref. 20 and in the Supplementary Information.

Received 6 October 2003; accepted 15 January 2004; doi:10.1038/nature02344.

- Goreau, T. F., Goreau, N. I. & Goreau, T. J. Corals and coral reefs. *Sci. Am.* **241**, 124–135 (1979).
- Crossland, C., Barnes, D. & Borowitzka, M. Diurnal lipid and mucus production in the staghorn coral *Acropora acuminata*. *Mar. Biol.* **60**, 81–90 (1980).
- Davies, P. S. The role of zooxanthellae in the nutritional energy requirements of *Pocillopora eydouxi*. *Coral Reefs* **2**, 181–186 (1984).
- Muscattine, L., McCloskey, L. R. & Marian, R. E. Estimating the daily contribution of carbon from zooxanthellae to coral animal respiration. *Limnol. Oceanogr.* **26**, 601–611 (1981).
- Meikle, P., Richards, G. & Yellowlees, D. Structural investigations on the mucus from 6 species of coral. *Mar. Biol.* **99**, 187–193 (1988).
- Ducklow, H. & Mitchell, R. Bacterial populations and adaptations in the mucus layers on living corals. *Limnol. Oceanogr.* **24**, 715–725 (1979).
- Krupp, D. A. Mucus production by corals exposed during an extreme low tide. *Pacif. Sci.* **38**, 1–11 (1984).
- Schuhmacher, H. *Proc. 3rd Int. Coral Reef Symp. Miami, Florida* 503–509 (University of Miami, Florida, 1977).
- Coles, S. & Strathman, R. Observations on coral mucus flocs and their potential trophic significance. *Limnol. Oceanogr.* **18**, 673–678 (1973).
- Marshall, M. Observations on organic aggregates in the vicinity of coral reefs. *Mar. Biol.* **2**, 50–55 (1968).
- Johannes, R. Ecology of organic aggregates in the vicinity of a coral reef. *Limnol. Oceanogr.* **12**, 189–195 (1967).
- Romaine, S., Tambutte, E., Allemand, D. & Gattuso, J. P. Photosynthesis, respiration and calcification of a zooxanthellate scleractinian coral under submerged and exposed conditions. *Mar. Biol.* **129**, 175–182 (1997).
- Moriarty, D. J. W., Pollard, P. C. & Hunt, W. G. Temporal and spatial variation in bacterial production in the water column over a coral reef. *Mar. Biol.* **85**, 285–292 (1985).
- Ferrier-Pages, C., Leclercq, N., Jaubert, J. & Pelegri, S. P. Enhancement of pico- and nanoplankton growth by coral exudates. *Aquat. Microb. Ecol.* **21**, 203–209 (2000).
- Vacelet, E. & Thomassin, B. Microbial utilization of coral mucus in long term *in situ* incubation over a coral reef. *Hydrobiologia* **211**, 19–32 (1991).
- Shanks, A. L. & Edmondson, E. W. Laboratory-made artificial marine snow—a biological model of the real thing. *Mar. Biol.* **101**, 463–470 (1989).
- Riedl, R. J., Huang, N. & Machan, R. The subtidal pump: a mechanism of interstitial water exchange by wave action. *Mar. Biol.* **13**, 210–221 (1972).
- Parnell, K. E. Water movement within a fringing reef flat, Orpheus Island, North Queensland, Australia. *Coral Reefs* **5**, 1–6 (1986).

- Oberdorfer, J. A. & Buddemeier, R. W. Coral-reef hydrology: field studies of water movement within a barrier reef. *Coral Reefs* **5**, 7–12 (1986).
- Wild, C. et al. Degradation and mineralization of coral mucus in reef environments. *Mar. Ecol. Prog. Ser.* **267**, 159–171 (2004).
- Marsden, J. & Meeuwij, J. Preferences of planktotrophic larvae of the tropical serpulid *Spirobranchus giganteus* (Pallas) for exudates of corals from a Barbados reef. *J. Exp. Mar. Biol. Ecol.* **137**, 97–104 (1990).
- Benson, A. & Muscatine, L. Wax in coral mucus—energy transfer from corals to reef fishes. *Limnol. Oceanogr.* **19**, 810–814 (1974).
- Richman, S., Loya, Y. & Slobodkin, L. Rate of mucus production by corals and its assimilation by the coral reef copepod *Acartia negligens*. *Limnol. Oceanogr.* **20**, 918–923 (1975).
- Wilkinson, C. R. Microbial ecology on a coral reef. *Search* **18**, 31–33 (1987).
- Richter, C., Wunsch, M., Rasheed, M., Kotter, I. & Badran, M. I. Endoscopic exploration of Red Sea coral reefs reveals dense populations of cavity-dwelling sponges. *Nature* **413**, 726–730 (2001).
- Herndl, G. J. & Velimirov, B. Microheterotrophic utilization of mucus released by the Mediterranean coral *Cladocora cespitosa*. *Mar. Biol.* **90**, 363–369 (1986).
- Stimson, J. & Kinzie, R. A. The temporal pattern and rate of release of zooxanthellae from the reef coral *Pocillopora damicornis* (Linnaeus) under nitrogen-enrichment and control conditions. *J. Exp. Mar. Biol. Ecol.* **153**, 63–74 (1991).
- Huetzel, M. & Gust, G. Solute release mechanisms from confined sediment cores in stirred benthic chambers and flume flows. *Mar. Ecol. Prog. Ser.* **82**, 187–197 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Alisch, S. Menger, H. Woyt, S. Gonelli and L. Hönemann for experimental assistance and for help with chemical analyses; O. Hoegh-Guldberg, R. Johnstone, T. Upton, R. Forbes and other staff members of Heron Island Research Station for logistical assistance; and P. Cook, C. Richter, R. Tollrian and H. Zech for improving the manuscript. All sample collections and *in situ* experiments were carried out under permits issued by the Great Barrier Reef Marine Park Authority. The Max Planck Society, Germany, funded this research.

Authors' contributions M.H. and C.W. conceptually designed and coordinated all experimental work, made most of the measurements, and wrote the manuscript. A.K. quantified coral distribution on Heron Island. S.K. and M.R. helped with chamber experiments and water analyses. B.B.J. contributed with ideas and advice to the significant improvement of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.W. (cwild@mpi-bremen.de).

Optimal traffic organization in ants under crowded conditions

Audrey Dussutour^{1,2}, Vincent Fourcassie¹, Dirk Helbing³ & Jean-Louis Deneubourg²

¹Centre de Recherches sur la Cognition Animale, UMR CNRS 5169, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 4, France

²Centre d'Études des Phénomènes Non-linéaires et des Systèmes Complexe, Université Libre de Bruxelles, CP231, Boulevard du Triomphe, 1050 Bruxelles, Belgium

³Institute for Economics and Traffic, Dresden University of Technology, D-01062 Dresden, Germany

Efficient transportation, a hot topic in nonlinear science¹, is essential for modern societies and the survival of biological species. Biological evolution has generated a rich variety of successful solutions², which have inspired engineers to design optimized artificial systems^{3,4}. Foraging ants, for example, form attractive trails that support the exploitation of initially unknown food sources in almost the minimum possible time^{5,6}. However, can this strategy cope with bottleneck situations, when interactions cause delays that reduce the overall flow? Here, we present an experimental study of ants confronted with two alternative routes. We find that pheromone-based attraction generates one trail at low densities, whereas at a high level of crowding, another trail is established before traffic volume is affected, which guarantees that an optimal rate of food return is maintained. This bifurcation phenomenon is explained by a

nonlinear modelling approach. Surprisingly, the underlying mechanism is based on inhibitory interactions. It points to capacity reserves, a limitation of the density-induced speed reduction, and a sufficient pheromone concentration for reliable trail perception. The balancing mechanism between cohesive and dispersive forces appears to be generic in natural, urban and transportation systems.

Animals living in groups^{7,8} often display collective movement along well-defined lanes or trails^{9–13}. This behaviour emerges through self-organization resulting from the action of individuals on the environment^{14–16}. In ants^{17,18}, for example, mass recruitment allows a colony to make adaptive choices based solely on information collected at the local level by individual workers. When a scout ant discovers a food source, it lays an odour trail on its way back to the nest. Recruited ants use the trail to find the food source and reinforce it in turn on their way back to the nest. Without significant crowding, mass recruitment generally leads to the use of only one trail¹⁹, because small initial differences in pheromone concentration between the trails are amplified as greater numbers of ants choose the trail that was initially slightly stronger, and hence reinforce it. This, together with travel time, ant numbers and ant individual behaviours, also explains why ants use the shortest among several paths leading to the same food source^{19,20}.

Here, we investigate the regulation of traffic flow during mass recruitment in the black garden ant *Lasius niger*. Ants were forced to move on a diamond-shaped bridge that formed two branches between their nest and a food source (Fig. 1). We studied to what extent the traffic on the bridge remains asymmetrical (and therefore limited by the capacity of one branch) when an increased level of crowding is induced by using branches of reduced width ($w = 10.0, 6.0, 3.0$ and 1.5 mm). The temporal evolution of the flow of ants on the bridge shown in Fig. 2 is typical of a trail recruitment process²¹. Surprisingly, the recruitment dynamics was not influenced by the branch width w and the traffic volumes were the same (analysis of variance, ANOVA, $F_{3,44} = 0.500$, $P = 0.690$). However, for $w = 10$ mm, the majority of ants used the same branch (Fig. 3a), whereas for $w \leq 6$ mm most experiments led to symmetrical traffic ($\chi^2 = 1.686$, d.f. = 2, $P = 0.430$). This suggests that there is a transition from asymmetrical to symmetrical traffic between 10.0 and 6.0 mm ($\chi^2 = 12.667$, d.f. = 3, $P = 0.005$).

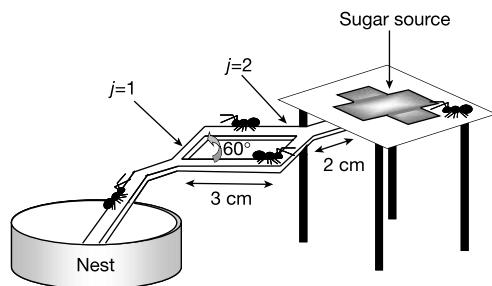


Figure 1 Experimental set-up. Five queenless colonies of *L. niger* (Hymenoptera, Formicidae), each containing 500 workers, were used. During an experiment, ants had access from the nest to a source of food (2 ml of 1 M sucrose solution) placed at the end of a diamond-shaped bridge. The solution was spread over a large surface so that all ants arriving at the end of the bridge could have access to the food. $j = 1$ and $j = 2$ indicate the choice points for branch selection. The whole set-up was surrounded by white curtains to prevent any bias in branch choice caused by the use of visual cues. The colonies were starved for four days before each experiment. The traffic on the two-branch bridge was recorded by a video camera for one hour. Nestbound and outbound ants were then counted and aggregated over 1-min intervals. Counting began as soon as the first ant climbed the bridge. Four different bridge widths (w) were used. This set-up mimics many natural situations in which the geometry of the trails is influenced by physical constraints of the environment, such as the diameter of underground galleries or of the branches in the vegetation along which the ants are moving.

We have identified the mechanism for this ‘symmetry-restoring transition’ using experiments, analytical calculations and Monte Carlo simulations. Figure 3b shows that the proportion of experiments producing symmetrical flows on narrow bridges increases with the total number of ants crossing the bridge ($\chi^2 = 4.6$, d.f. = 2, $P = 0.04$). Moreover, Fig. 3c shows that both branches were equally used by the opposite flow directions. Therefore, in contrast to army ants⁶ and pedestrians¹, separation of the opposite flow directions was not found. This could be explained by a high level of disturbance, such as a large variation in the speeds of ants²².

The flow of ants over the bridge can be understood analytically. The concentration of the trail pheromone C_{ij} on branch i ($i = 1, 2$) immediately behind each choice point j ($j = 1, 2$) changes in time t according to:

$$dC_{ij}/dt = q\Phi_{ij}(t) + q\Phi_{ij'}(t - \tau) - \nu C_{ij}(t) \text{ with } j' = 3 - j \quad (1)$$

where $\Phi_{i1}(t)$ represents the overall flow of foragers from the nest to the food source that choose branch i behind the choice point 1, $\Phi_{i2}(t)$ is the opposite flow on branch i behind the other choice point $j' = 3 - j = 2$, τ is the average time required for an ant to get from one choice point to the other, q is the quantity of pheromone laid on the trail, and ν is the decay rate of the pheromone. Moreover, if the density is low enough:

$$\Phi_{ij}(t) = \phi_j(t)F_{ij}(t) \quad (2)$$

where ϕ_1 is the outbound flow of foragers from the nest to the food source and ϕ_2 is the opposite, nestbound flow. The function F_{ij} describes the relative attractiveness of the trail on branch i at choice point j . For *L. niger* it is given by¹⁹:

$$F_{ij} = \frac{(k + C_{ij})^2}{(k + C_{1j})^2 + (k + C_{2j})^2} = 1 - F_{i'j} \text{ with } i' = 3 - i \quad (3)$$

where k denotes a concentration threshold beyond which the pheromone-based choice of a trail begins to be effective.

Equation (2) describes the flow dynamics without interactions between ants. However, when traffic was dense, we observed that ants that had just started to move on one branch were often pushed to the other branch, having collided frontally with another ant coming from the opposite direction. This behaviour will turn out to be essential in generating symmetrical traffic on narrow bridges.

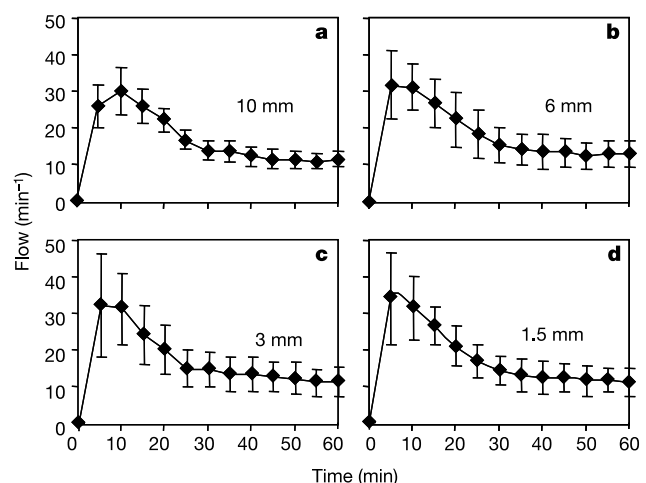


Figure 2 Average number of ants per minute crossing the two branches of the bridge within intervals of five minutes. The observed traffic volumes $2\phi(t)$ and flow dynamics agreed for the four different branch widths, that is, the bottleneck for smaller branches was compensated for by use of both branches. **a**, $w = 10$ mm; **b**, $w = 6$ mm; **c**, $w = 3$ mm; **d**, $w = 1.5$ mm. $N = 12$ experiments were carried out for each bridge width. Error bars indicate the standard deviation.

Whereas pushing was practically never observed on a 10-mm bridge, on narrow bridges the frequency of pushing events was high and proportional to the flow (Fig. 3d).

When pushing is taken into account, the overall flow of ants arriving at choice point j and choosing branch i can be expressed by the following formula:

$$\Phi_{ij}(t) = \phi_j(t)F_{ij}(t)[1 - \gamma a \Phi_{ij'}(t - \tau)/w] + \phi_j(t)F_{i'j}(t)\gamma a \Phi_{ij'}(t - \tau)/w \quad (4)$$

The first term on the right-hand side of equation (4) represents the flow of ants engaged on branch i (that is, $\phi_j(t)F_{ij}(t)$), diminished by the flow of ants pushed towards the other branch i' by ants arriving from the opposite direction. $a\Phi_{ij'}(t - \tau)/w$ is the proportion of ants decelerated by interactions. The factor a is proportional to the interaction time period and the lateral width of ants. Up to the symmetry-restoring transition it can be considered approximately constant. $\gamma \approx 0.57$ denotes the probability of being pushed in the case of an encounter (Fig. 3d). The second term on the

right-hand side of equation (4) represents the flow of ants that were engaged on branch i' and were pushed towards branch i .

The stationary solutions of this model are defined by the conditions $dC_{ij}/dt = 0$, $F_{ij}(t - \tau) = F_{ij}(t) = F_{ij}$, $\Phi_{ij}(t - \tau) = \Phi_{ij}(t) = \Phi_{ij}$ and $\phi_j(t - \tau) = \phi_j(t) = \phi_j' = \phi$ (because the nest-bound flow and the outbound flow should be equal). These imply:

$$C_{ij} = q(\Phi_{ij} + \Phi_{ij'})/\nu = C_{ij'}, \quad F_{ij} = F_{ij'} = 1 - F_{i'j} = 1 - F_{i'j'}$$

and

$$C_{ij} = C_{ij'} = \frac{q\phi}{\nu} + D, \quad C_{i'j} = C_{i'j'} = \frac{q\phi}{\nu} - D \quad (5)$$

with

$$D = 0 \quad (6)$$

or

$$D^2 = \frac{q^2\phi^2}{\nu^2} - \frac{k^2 + \gamma a \phi(k + 2q\phi/\nu)^2/w}{1 + \gamma a \phi/w} \quad (7)$$

We can distinguish two cases. Case (1): when $D^2 > 0$, the stable

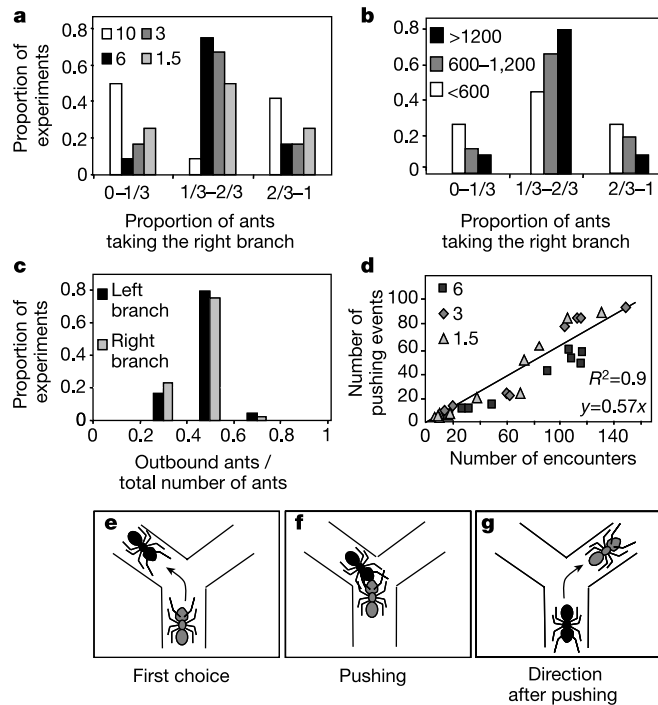


Figure 3 Experimental results. **a**, Experimental frequency distributions of the proportion F_{1j} of ants using the right branch for different branch widths. We considered traffic to be asymmetrical when more than 2/3 of the cumulated traffic of ants at the end of the experiment had used a single branch. All experiments have been pooled. **b**, Experimental frequency distributions similar to **a**, but for different total numbers of ants crossing the bridge. All experiments on bridges of $w = 1.5, 3.0$ and 6.0 mm have been pooled. **c**, Proportion of outbound ants on the left and right branches of the bridge divided by the total number of ants that have passed the bridge at the end of the experiment. The results contradict opposite one-way flows on both branches. **d**, Number of pushing events as a function of the total number of ant encounters on the initial part of each branch of the bridge. For each branch width, pushing events were evaluated at both choice points, for outbound and nestbound ants, during the first ten and last ten minutes of a random sample of two experiments characterized by symmetrical and two experiments characterized by asymmetrical traffic. This yielded a total of 2 branches $\times$ 4 experiments $\times$ 2 time intervals = 16 points per branch width. When the number of encounters was too low (≤ 3) the points were not taken into account in the regression. The slope of the linear regression is equal to 0.571 ± 0.057 (confidence interval, $CI_{95\%}$). This corresponds to the probability γ that an ant will be pushed and redirected towards the other branch after encountering another ant coming from the opposite direction (see **e-g**).

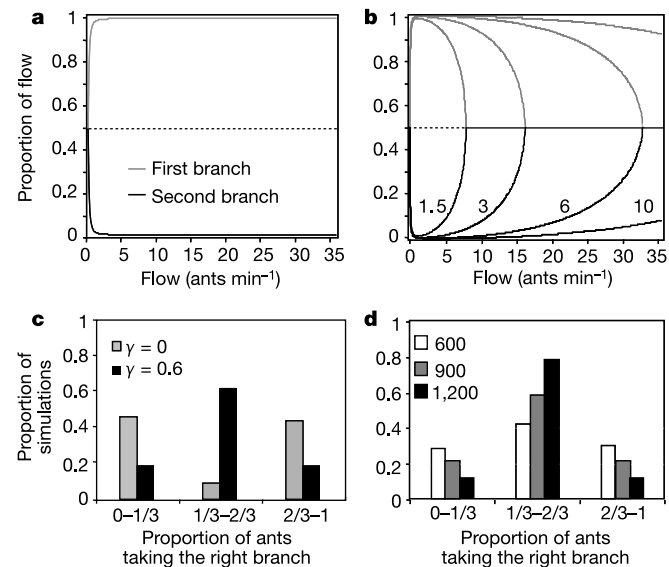


Figure 4 Analytical and Monte Carlo simulation results. **a, b**, Analytical results for the parameter values $q = 1$, $k = 6$, $\nu = 1/40 \text{ min}^{-1}$, and $a = 0.1 \text{ mm min}^{-1}$, which have been adjusted to experimental measurements. The curves show the proportion F_{ij} of the overall flow ϕ on each branch in the stationary state. **a**, In the absence of pushing ($\gamma = 0$) the model predicts asymmetrical traffic above very low values of the overall flow of ants, independently of the traffic volume. **b**, When the proportion of pushed ants is high ($\gamma = 0.6$), traffic is asymmetrical for moderate values of the overall flow of ants, but stable symmetrical traffic with $F_{ij} = 0.5$ is expected above a critical value of traffic flow, which is an increasing function of branch width. The symmetry-restoring transition from asymmetrical to symmetrical traffic flow corresponds to an inverse pitchfork bifurcation. **c, d**, Results of Monte Carlo simulations. At time $t = 0$, the pheromone concentration and the number of ants on each branch are set to zero. Ants arrive at choice point j at a rate $\phi_j(t) = \phi$. The probability of choosing the right or left branch at a choice point is governed by the choice function F_{ij} (see equation (3)). However, as soon as an ant has engaged on a branch, it may be pushed to the other branch with probability γ by an ant moving in the opposite direction coming from the second choice point. The pushed ant then continues its course on the alternative branch and lays a trail. For each value of γ , the simulations are averaged over 1,000 runs. The graphs show the relative frequency of simulations in which a certain proportion of traffic was supported by the right branch. **c**, Assuming a pushing probability equal to zero ($\gamma = 0$), most simulations ended with asymmetrical traffic. However, when we used the experimental value $\gamma = 0.6$, most simulations for narrow branches resulted in symmetrical traffic. **d**, The proportion of simulations with $\gamma = 0.6$ in which asymmetrical traffic emerged is a function of the total number of ants crossing the bridge, just as in our experiments (compare Fig. 3b).

stationary solution corresponds to asymmetrical traffic with $C_{ij} = \frac{q\phi}{v} \pm \sqrt{D^2}$ and $C_{ij'} = \frac{q\phi}{v} \mp \sqrt{D^2}$. If $\gamma = 0$, this situation is given for $q\phi/v > k$. If $q\phi/v \leq k$, that is, if the pheromone concentration is too low, the ants choose both branches at random, corresponding to a symmetrical distribution (Fig. 4a). Case (2): when pushing is taken into account with $\gamma > 0$, the organization of traffic changes considerably: the asymmetrical solution can no longer be established as soon as $D^2 < 0$ for large traffic volumes ϕ (Fig. 4b). In this case, symmetrical traffic is expected with $C_{ij} = q\phi/v$ and $\Phi_{ij} = \phi/2$ for both branches i and choice points j . Therefore, high traffic volumes can be maintained and none of the branches is preferred in spite of the competitive effect due to the accumulation of pheromone on both branches. Moreover, the model implies that the outbound flow Φ_{i1} and the nestbound flow Φ_{i2} are equal, indicating that one-way flows are not required to maintain a high traffic volume²³. These analytical results have been confirmed by Monte Carlo simulations (see Fig. 4c, d). Our simulations have also demonstrated that if pushed ants, instead of moving to the other branch, made a U-turn and returned in the direction they had been coming from, the overall flow of ants crossing the bridge was affected by the branch width and no shift to symmetrical traffic was observed.

The traffic organization in ants can be called optimal. The overall flow on branch i behind choice point j is given by $\Phi_{ij} = w\rho_{ij}V_{ij} \leq \phi$, where ρ_{ij} denotes the density of ants. Their average speed is theoretically estimated by $V_{ij} \approx V_m(1 - a\Phi_{ij}/w)$, where V_m denotes the average maximum speed and $a\Phi_{ij}/w$ is again the proportion of decelerated ants. For the symmetry-restoring transition with $D^2 = 0$, equation (7) requires $a\gamma\phi/w < 1/3$, which implies $V_{ij} > V_m[1 - 1/(3\gamma)] \approx 0.42V_m$. However, according to the empirical speed–density relation by Burd *et al.*²³, the maximum flow (the capacity) is only reached at the smaller speed $V_{ij} = V_m[1 - 1/(n+1)] \approx 0.39V_m$. Although the empirical value $n \approx 0.64$ was determined for another ant species, the values for *L. niger* should be comparable. This shows that the symmetry-restoring transition occurs before the maximum flow is reached. The strict inequality also implies capacity reserves and a limitation of the density-related speed reduction. A marginal reduction in speed, however, would not favour symmetrical traffic because of the benefits of using a single trail. First, a more concentrated trail provides a better orientation guidance and a stronger arousal stimulus²⁴. Second, a higher density of ants enhances information exchange and supports a better group defence²⁵.

We have demonstrated the surprising functionality of collisions among ants to keep up the desired flow level by generation of symmetrical traffic. Pushing behaviour may be considered an optimal behaviour to maintain a high rate of food return to the nest. It would not be required in most models of ideal free distribution (IDF)²⁶, as they neglect effects of inter-attraction. The balancing between cohesive and dispersive forces avoids a dysfunctional degree of aggregation and supports an optimal accessibility of space at minimal costs allowing an efficient construction, maintenance and use of infrastructures. This mechanism appears to be generic in nature, in particular in group-living organisms. For example, inhibitory interactions at overcrowded building sites in termites allow a smooth growth of the nest structure²⁷. In the development of urban agglomerations such interactions help to maintain the coexistence of distinct cities²⁸, and in vehicle traffic they determine the choice of longer, less congested routes. The mechanism also suggests algorithms for the routing of data traffic in networks. □

Received 11 December 2003; accepted 13 January 2004; doi:10.1038/nature02345.

1. Helbing, D. Traffic and related self-driven many-particle systems. *Rev. Mod. Phys.* **73**, 1067–1141 (2001).
2. Camazine, S., *et al.* *Self-organized Biological Systems* (Princeton Univ. Press, Princeton, 2001).
3. Bonabeau, E., Dorigo, M. & Theraulaz, G. Inspiration for optimization from social insect behaviour. *Nature* **406**, 39–42 (2000).
4. Bonabeau, E., Dorigo, M. & Theraulaz, G. *Swarm Intelligence: From Natural to Artificial Systems* (Oxford Univ. Press, New York, 1999).

5. Burd, M. & Aranwela, N. Head-on encounter rates and walking speed of foragers in leaf-cutting ant traffic. *Insectes Soc.* **50**, 3–8 (2003).
6. Couzin, I. D. & Franks, N. R. Self-organized lane formation and optimized traffic flow in army ants. *Proc. R. Soc. Lond. B* **270**, 139–146 (2003).
7. Parrish, J. K. & Hamner, W. K. *Animal Groups in Three Dimensions* (Cambridge Univ. Press, Cambridge, UK, 1997).
8. Krause, J. & Ruxton, D. *Living in Groups* (Oxford Univ. Press, Oxford, UK, 2002).
9. Galef, B. G. Jr & Buckley, L. L. Using foraging trails by Norway rats. *Anim. Behav.* **51**, 765–771 (1996).
10. Fitzgerald, T. D. *The Tent Caterpillars* (Cornell Univ. Press, Ithaca, NY, 1995).
11. Erlandson, J. & Kostylev, V. Trail following, speed and fractal dimension of movement in a marine prosobranch, *Littorina littorea*, during a mating and a non-mating season. *Mar. Biol.* **122**, 87–94 (1995).
12. Traniello, J. F. A. & Robson, S. K. in *Chemical Ecology of Insects* (eds Cardé, R. T. & Bell, W.) 241–286 (Chapman & Hall, New York, 1995).
13. Miura, T. & Matusmoto, T. Open-air litter foraging in the nasute termite *Longipeditermes longipes* (Isoptera: Termitidae). *J. Insect Behav.* **11**, 179–189 (1998).
14. Helbing, D., Keltsch, J. & Molnár, P. Modelling the evolution of human trail systems. *Nature* **388**, 47–50 (1997).
15. Edelstein-Keshet, L. W., Ermentrout, J. & Bard, G. Trail following in ants: individual properties determine population behaviour. *Behav. Ecol. Sociobiol.* **36**, 119–133 (1995).
16. Schweitzer, F. *Brownian Agents and Active Particles. Collective Dynamics in the Natural and Social Sciences* (Springer, Berlin, 2003).
17. Wilson, E. O. Chemical communication among workers of the fire ant *Solenopsis saevissima* (Fr. Smith). 1. The organization of mass-foraging. *Anim. Behav.* **10**, 134–147 (1962).
18. Hölldobler, B. & Wilson, E. O. *The Ants* 265–279 (The Belknap Press of Harvard Univ. Press, Cambridge, MA, 1990).
19. Beckers, R., Deneubourg, J. L. & Goss, S. Trails and U-turns in the selection of a path by the ant *Lasius niger*. *J. Theor. Biol.* **159**, 397–415 (1992).
20. Goss, S., Aron, S., Deneubourg, J. L. & Pasteels, J. M. Self-organized shortcuts in the Argentine ant. *Naturwissenschaften* **76**, 579–581 (1989).
21. Pasteels, J. M., Deneubourg, J. L. & Goss, S. in *From Individual to Collective Behaviour in Social Insects* (eds Pasteels, J. M. & Deneubourg, J. L.) *Experientia* suppl. Vol. 54, 155–175 (Birkhäuser, Basel, 1987).
22. Helbing, D. & Platkowski, T. Drift- or fluctuation-induced ordering and self-organization in driven many-particle systems. *Europhys. Lett.* **60**, 227–233 (2002).
23. Burd, M. *et al.* Traffic dynamics of the leaf-cutting ant *Atta cephalotes*. *Am. Nat.* **159**, 283–293 (2002).
24. Calenbuhr, V. *et al.* A model for osmotropotactic orientation. II. *J. Theor. Biol.* **158**, 395–407 (1992).
25. Feener, D. H. & Moss, K. A. G. Defense against parasites by hitchhikers in leaf-cutting ants: a quantitative assessment. *Behav. Ecol. Sociobiol.* **49**, 348–356 (1990).
26. Fretwell, S. D. & Lucas, H. L. On territorial behavior and other factors influencing habitat distribution in birds. I. Theoretical development. *Acta Biotheor.* **19**, 16–36 (1970).
27. O'Toole, D. V., Robinson, P. A. & Myerscough, M. R. Self-organized criticality in termite architecture: a role for crowding in ensuring ordered nest expansion. *J. Theor. Biol.* **198**, 305–327 (1999).
28. Fujita, M., Krugman, P. & Venables, A. *The Spatial Economy—Cities, Regions, and International Trade* (MIT Press, Cambridge, MA, 1999).

Acknowledgements We thank G. Theraulaz and all the members of his team 'Ethology and modelization of collective behaviours' for discussions and comments on the manuscript. We also thank A. Schadschneider for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.D. (dussutou@cict.fr).

Perceiving distance accurately by a directional process of integrating ground information

Bing Wu¹, Teng Leng Ooi² & Zijiang J. He¹

¹Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292, USA

²Department of Basic Sciences, Pennsylvania College of Optometry, 8360 Old York Road, Elkins Park, Pennsylvania 19027, USA

By itself, the absolute distance of an object cannot be accurately judged beyond 2–3 m (refs 1–3). Yet, when it is viewed with reference to a flat terrain, humans accurately judge the absolute distance of the object up to 20 m, an ability that is important for various actions^{4–8}. Here we provide evidence that this is accomplished by integrating local patches of ground information into a global surface reference frame. We first show that restricting an

observer's visual field of view to the local ground area around the target leads to distance underestimation, indicating that a relatively wide expanse of the ground surface is required for accurate distance judgement. Second, as proof of surface integration, we show that even with the restricted view, the observer can accurately judge absolute distance by scanning local patches of the ground surface, bit by bit, from near to far, but not in the reverse direction. This finding also reveals that the surface integration process uses the near-ground-surface information as a foundation for surface representation, and extrapolation to the far ground surface around the target for accurate absolute distance computation.

Experiment 1 examined how the visual field of view affects judged absolute distance, to show that extended ground surface is required for accurate space vision. The observer wore a pair of opaque goggles with a monocular rectangular aperture that limited the field of view to the flat grass field surrounding the target. During the trial, the observer kept his head still, judged the target distance, then closed his eyes and walked blindly to traverse the remembered target distance. Figure 1a plots the averaged distances walked, reflecting the judged absolute distances in three aperture-size conditions. When the field of view was $38.6^\circ \times 39.5^\circ$, the walked distances were accurate and similar to the full-view condition⁹ (control) ($F(1,7) = 0.012$, $P = 0.916$; $F(3,21) = 0.370$, $P = 0.775$). Significantly, distance was underestimated when the field of view was reduced to $21.2^\circ \times 21.2^\circ$ ($F(1,7) = 18.267$, $P < 0.005$; $F(3,21) = 1.651$, $P = 0.208$) and $13.9^\circ \times 13.5^\circ$ ($F(1,7) = 60.857$, $P < 0.001$; $F(3,21) = 9.702$, $P < 0.001$). These findings reveal that the size of the visible field is significant in absolute distance judgement^{10,11}; that is, large-scale ground-surface information (more than 21°) is essential.

To underscore the effect of field size, we next tested the prediction that a small visual field also exaggerates the error in relative length-in-depth judgement. Thus, in Experiment 2 the observers viewed an L-shaped target on the flat grass field in three aperture-size conditions (Fig. 2b). The adjustable length (Z) of the L-shaped target was positioned in depth, whereas its width was fixed ($W = 40.5$ cm) in the frontoparallel plane. The observer perceptually matched the Z and W , which allowed us to compute the perceptual aspect ratio, R , defined as W/Z (ref. 12). In this way, $R < 1$ indicates that the observer required a longer Z to match W ; that is, the relative length-in-depth of the L-shaped target is underestimated. The averaged results (Fig. 1b) show relative length-in-depth underestimation ($R < 1$) in all three aperture-size conditions. When the field size was $38.6^\circ \times 39.5^\circ$, the perceptual aspect ratios were similar to that in the full view (control) condition ($F(1,7) = 0.013$, $P = 0.914$; $F(2,14) = 0.077$, $P = 0.926$). Although such findings of depth compression are expected^{6,12,13}, depth compression becomes significantly larger than the full-view condition with the $21.2^\circ \times 21.2^\circ$ aperture ($F(1,7) = 17.001$, $P < 0.005$; $F(2,14) = 5.578$, $P < 0.025$), and larger still with the $13.9^\circ \times 13.5^\circ$ aperture ($F(1,7) = 60.017$, $P < 0.001$; $F(2,14) = 6.860$, $P < 0.01$). Clearly, decreasing the ground-surface view also adversely affects relative length-in-depth judgement (Fig. 1b), as it does absolute distance judgement (Fig. 1a).

Figure 2 points to a likely common basis for the findings of Experiments 1 and 2. When a patch of far ground surface (more than 2–3 m) is seen in isolation, the depth information consists mainly of texture gradient cues, which are insufficient for the visual system to code absolute distance accurately¹⁴. Moreover, the patch of far ground surface is perceived as slanted towards the vertical (the frontal tendency)^{15–17}. Consequently, when constrained to sampling only this far surface patch, the global visual representation of the ground (grey line in Fig. 2a) is based on this far patch with a perceived slant error of η degrees. This then compels the visual system to determine the target position on the ground on the basis of the slant surface representation.

Now that the target is represented as located on a surface with a slant error (η), whereas its angular declination (α) is correctly perceived³, its absolute distance, d , on the ground surface is obtained according to the trigonometric relationship (Fig. 2a)

$$d = DH / (H \cos \eta + D \sin \eta)$$

where D and H are the physical distance of the object on the ground and the observer's eye height, respectively. Thus, on the basis of this relationship, when the ground-surface representation is slanted upward (that is, $\eta > 0$), d becomes smaller than D , leading to absolute distance underestimation. Similarly, the perceptual aspect ratio (R) is described by

$$R = H / [H \cos \eta + (D + Z) \sin \eta]$$

where Z is the matched depth of the L-shaped target (Fig. 2b).

We applied the averaged absolute distance (d) and perceptual aspect ratios (R) data from Fig. 1a, b, respectively, into each equation above to derive the slant errors (η). As plotted in Fig. 2c, the slant errors in both absolute distance and relative length-in-depth judgements increase with decreasing field of view, even as the two sets of data differ⁷ because of the compression of depth in relative length-in-depth judgement ($\eta > 0$), whereas absolute distance judgement remains accurate ($\eta = 0$) in the full-view condition^{4–8,12,13}. Then, for a given test distance and visual field size, we related the derived slant errors (η) from relative length-in-depth to absolute distance judgements in a scatter plot (Fig. 2d). Clearly, a

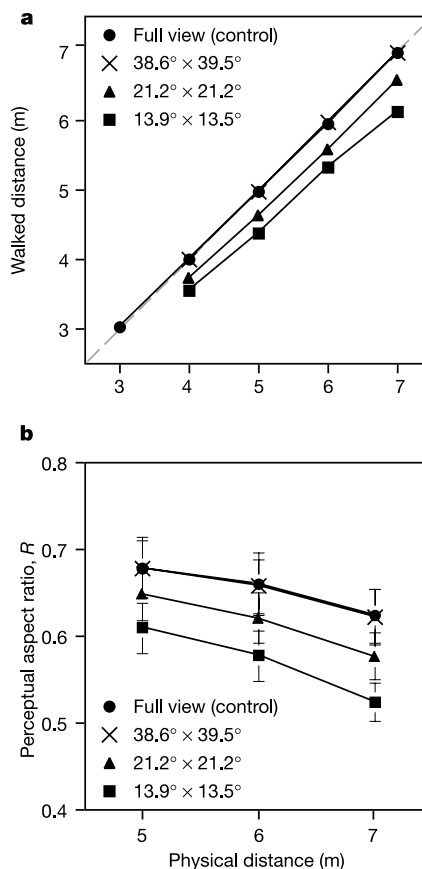


Figure 1 The effect of visual field size on judged distance. **a**, Absolute walked distance from the blind-walking task is plotted as a function of the physical distance. **b**, Judged relative length-in-depth from the perceptual matching task, defined as the perceptual aspect ratio (R) of the L-shaped target, is plotted as a function of the physical distance. Each symbol represents the averaged result of eight naive observers. For this and subsequent figures, if error bars are absent from selected symbols the standard errors are smaller than the sizes of the symbols.

significant correlation exists between both measures ($R^2 = 0.9018$, $P < 0.001$), indicating that a common piece of perceptual information⁶, namely the perceived surface slant due to the reduced view of the ground, might affect both types of space judgement. The perceived surface slant and its consequence for judged target distance (Fig. 2a, b) probably reflect the weighted contributions of the restricted texture information on the ground and the intrinsic bias of the visual system in representing the ground surface. (Measuring the implicit slant errors in the dark, in which the observer has no access to ground texture information, reveals the latter³.) It is reasonable to suggest that when the visual field size increases, the relative contribution of ground texture information to the ground-surface representation mechanism increases. This, then, improves the accuracy of representing the ground surface, and hence distance.

Evidently, distance judgements improve with larger field sizes, because the visual system is able to access more visual surface information for representing the ground. But which part of the ground holds more essential information? To answer this in Experiment 3, we compared absolute distance judgement when the visual field was delimited vertically and horizontally. The observers, without head movements, wore a pair of goggles with a vertical rectangular aperture that maintained the vertical field at 50.9° while restricting the horizontal view. Figure 3a shows that the averaged blind-walking performances were as accurate as in the full-view condition (control) for all three aperture-size conditions ($F(4,28) = 0.143$, $P = 0.965$; $F(12,84) = 0.863$, $P = 0.586$). The observers were also tested with horizontal apertures that maintained

the horizontal view at 57.7° while restricting the vertical view. The averaged results (Fig. 3b) show that absolute distance judgements with the 39.9° and 29.6° vertical field extents were as accurate as that in the full-view condition (39.9°: $F(1,7) = 1.473$, $P = 0.264$; $F(3,21) = 6.591$, $P < 0.001$; 29.6°: $F(1,7) = 4.151$, $P = 0.081$; $F(3,21) = 2.784$, $P = 0.066$); however, distances were underestimated when the vertical field extents were reduced to 21.1° ($F(1,7) = 27.220$, $P < 0.001$; $F(3,21) = 1.364$, $P = 0.291$) and 13.6° [$F(1,7) = 69.420$, $P < 0.001$; $F(3,21) = 1.207$, $P = 0.332$]. Our data indicate that for a target distance of 4 m and an average eye height of 1.67 m, the 21.1° vertical aperture blocked the view of the near ground surface up to about 2.55 m from the observer. Clearly, the near ground surface, which provides reliable depth information, is critical for representing the ground.

The findings of Experiment 3 underscore the necessity for the visual system to make use of not only the far ground around the target, but also the near-ground information^{18,19}. This suggests that representing the ground surface requires integrating local patches of ground surface from the observer to the target. Further, it can be surmised that the near-ground information serves as a foundation early on in the surface representation process, rather than late, to guarantee an accurate representation. Experiment 4 proves this by comparing absolute distance performances when observers were asked to scan the visual scene from the near to the far ground surface (near-to-far), in comparison with the reverse scanning direction (far-to-near). To scan the entire grass field while viewing through the aperture, the observer rotated his head once either upward from an initial downward head position (near-far-1 scanning), or

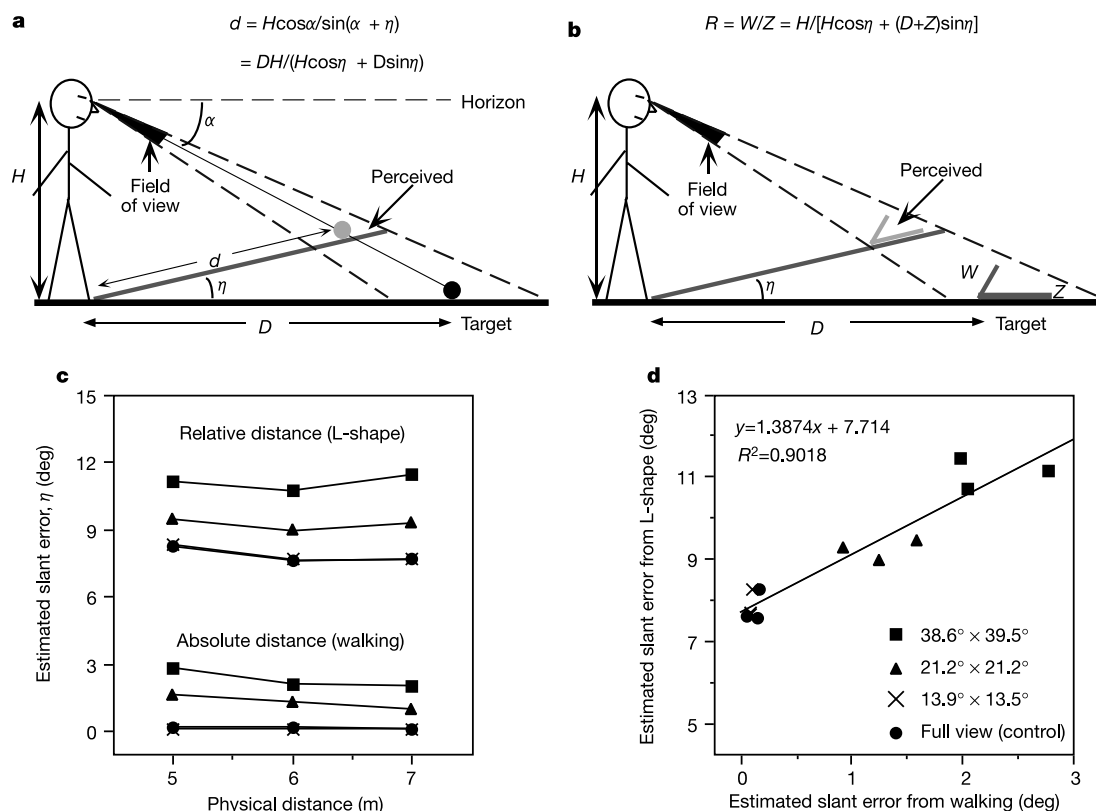


Figure 2 Profile illustrations of the physical and perceived ground surface through a limited field of view. **a**, **b**, A flat ground surface (black line) is visually represented as a slant surface (grey line) with a slant error of η . **a**, An observer with an eye height of H perceives a target (black disk) on the flat ground as one (grey disk) on the slant surface. As the perceived target's angular declination (α) remains correct, the perceived target distance, d , is determined by the equation above the figure. **b**, Similarly, the grey L-shaped target on the flat ground is perceived as one (lighter grey) on the slant surface.

The perceived relative length-in-depth is defined by the perceptual aspect ratio, $R = W/Z$, and is governed by the equation above the figure. **c**, Slant errors (η) are derived from the blind-walking and perceptual matching (L-shaped) data using the equations in **a** and **b**, respectively. The graph relates the estimated slant error to the physical distance. **d**, For each physical distance tested, the estimated slant errors from the relative length-in-depth (L-shaped) and absolute distance (blind-walking) tasks are related in the scatter plot.

downward from an initial straight-ahead position (far-near-1 scanning). Figure 4 shows the averaged blind-walking results from two different aperture sizes, where open triangles depict performance for near-to-far, and filled triangles show performance for far-to-near. Clearly, the latter was underestimated whereas the former remained as accurate as in the full-view condition ($57.7^\circ \times 13.6^\circ$ aperture: $F(1,7) = 19.759$, $P < 0.005$; $F(3,21) = 12.355$, $P < 0.001$; $57.7^\circ \times 21.1^\circ$ aperture: $F(1,7) = 41.429$, $P < 0.001$; $F(3,21) = 14.362$, $P < 0.001$). These results indicate that accurate surface representation is obtained through surface integration, and only when the near ground surface is accessed before the far.

However, perhaps for lack of iteration, the visual system was unable to form an accurate surface representation with only one chance to scan from the far ground to near. We therefore asked the observers to scan twice. In a far-near-2 condition, the observer rotated his head downwards to scan from the far ground to near, then pulled a blindfold over his goggles and repositioned his head (straight), removed the blindfold, and rotated his head downward to scan from far to near again, before performing the blind-walking task. For completeness, we included a near-far-2 condition in which scanning from near to far was performed twice. The filled squares in Fig. 4 represent the far-to-near scanning performances, which were underestimated in comparison with the near-to-far scanning performances (open squares) ($57.7^\circ \times 13.6^\circ$ aperture: $F(1,7) = 18.642$,

$P < 0.005$; $F(3,21) = 11.804$, $P < 0.001$; $57.7^\circ \times 21.1^\circ$ aperture: $F(1,7) = 22.087$, $P < 0.005$; $F(3,21) = 6.759$, $P < 0.005$). Altogether, finding that scanning twice in the 'wrong' direction (far-near-2) did not improve performance suggests strongly that the visual representation of the near ground serves as an anchor, or foundation, early on in the surface representation process to build an accurate representation.

We conducted a control experiment to negate the possibility that the asymmetric performance in Experiment 4 was due to the motor system responsible for head movement. We compared blind-walking performances in judging the location (distance and height) of a light target² in three viewing conditions (free head motion, near-to-far head motion, and far-to-near head motion) in the dark, where other visual cues are unavailable (see Methods). The observers performed similarly in all three conditions, indicating that without a visible ground surface the scanning direction is irrelevant (distance: $F(2,14) = 0.764$, $P = 0.484$; $F(6,42) = 1.664$, $P = 0.154$; height: $F(2,14) = 2.613$, $P = 0.109$; $F(6,42) = 0.882$, $P = 0.517$). The asymmetry in Experiment 4 therefore reveals the process of representing visual surface information.

In summary, our findings reveal that for accurate distance judgement the human visual system relies on a surface integration process to form a global ground-surface representation, which is used as a reference frame for coding distances^{3,8,14,19,20}. The surface integration process has a directional dictate, from near to far, for it to use the reliable visual depth information on the near ground surface. Furthermore, having the observer sample local patches of texture information by scanning reminds us of a parallel in nature.

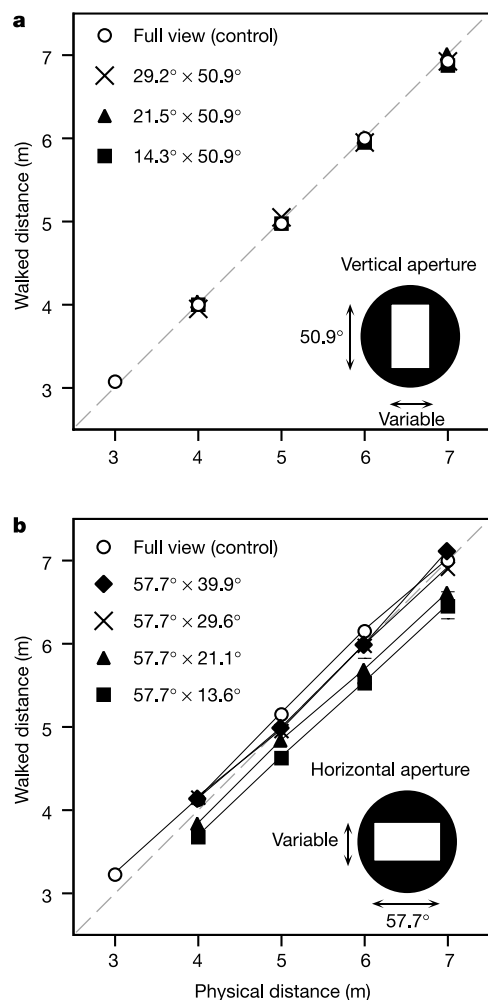


Figure 3 The critical ground surface for accurate distance judgement. Comparison of the impact of vertical (**a**) and horizontal (**b**) rectangular apertures on the judged absolute distance measured with the blind-walking task. Each symbol represents the averaged result of eight naive observers.

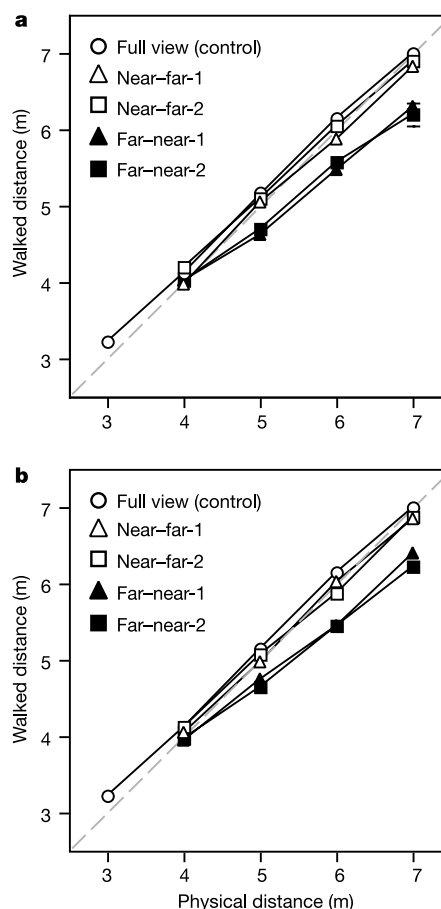


Figure 4 Scanning direction affects judged absolute distance. **a**, **b**, Data from the $57.7^\circ \times 13.6^\circ$ (**a**) and $57.7^\circ \times 21.1^\circ$ (**b**) aperture-size experiments. Each symbol represents the averaged result of eight naive observers.

In particular, lower animals with few photoreceptor cells—and therefore very limited visual fields—use scanning to detect objects and to learn or recognize landmarks beyond their visual field coverage^{21,22}. In humans, scanning local information with the fovea during surface integration also ensures that the information sampled has better spatial resolution than that afforded by the peripheral retina²². □

Methods

Observers

Ten naive observers with self-reported normal vision and informed consent participated in the various experiments. All performed the experiments with their motor dominant, right eye. To limit the monocular field of view to a preselected size, a pair of clear safety goggles was painted black all over except for a rectangular area (aperture) in front of the right eye with the appropriate field extent. For all aperture sizes used, the observers were unable to see their body and feet without rotating their heads. The goggles were worn throughout the experiments except during the full-view (control) condition, in which the observers wore an opaque patch over the left eye without the goggles. They were not given any feedback about their performances.

Judging absolute distance by using the blind-walking task

The observer previewed an orange-red disk target (7.62 cm in diameter and 2.54 cm high) on a flat grass surface and judged its absolute distance. Then he pulled a blindfold over his eyes (goggles) and walked forward to traverse the remembered target distance. He stopped upon reaching the remembered target location and remained there to allow the experimenter to measure the walked distance. Thereafter, the experimenter led the observer, still in blindfold, back to the starting point to ready for a new trial. Four target distances (4, 5, 6 and 7 m) were tested in a randomized order. Each distance was tested twice and the average was taken as the final result.

Judging relative length-in-depth by using a perceptual matching task

Two white pipes, 3 cm in diameter, were used to construct the L-shaped target. The frontoparallel arm (*W*) of the L-shaped target was fixed at 40.5 cm, and the length of the arm in depth (*Z*) was adjustable. During a trial, the observer kept his head still and judged whether the width of arm-*Z* was equal to the length of arm-*W*. If not, the observer pulled a blindfold over his eyes (goggles) and instructed the experimenter to adjust the length of arm-*Z*. After the experimenter finished making the adjustment and had walked away from the L-shaped target, the observer was told to remove the blindfold and to compare the width and length of the L-shaped target again. This was repeated several times until the observer was satisfied that the two arms matched in width and length. The base of the L-shaped target to the observer defined the viewing distance, which was one of three distances (5, 6 and 7 m) that was measured in a randomized order. Each distance was tested twice and the average was taken as the final result.

Judging distance in the dark

The test target, a 0.23° internally illuminated red table tennis ball, was placed at one of four distances (2.50, 3.75, 5.00 and 6.25 m) either on the floor or 0.5 m above it. Trials with targets above the floor served as catch trials (one-third of the total trials). Other than the test target, the room was completely dark so that the observer could not access other visual cues. For each trial, the observer (*n* = 9) previewed the distance and height of the target with the dominant eye. After this, the target was extinguished and the observer walked blindly to the remembered target location. Upon reaching his destination, the observer gestured the perceived height of the remembered target with his hand. The walked distance and gestured height are taken as the judged distance and height of the target, respectively. This procedure was used for all three conditions tested (free head motion, near-to-far head motion and far-to-near head motion).

Data analysis

We applied the two-way analysis of variance with repeated-measures analysis to our data to obtain the *F* and *P* values indicated in the text. Following convention, the main effect and interaction are presented in order.

Received 12 August 2003; accepted 20 January 2004; doi:10.1038/nature02350.

- Gogel, W. C. & Tietz, J. D. A comparison of oculomotor and motion parallax cues of egocentric distance. *Vision Res.* **19**, 1161–1170 (1979).
- Philbeck, J. W. & Loomis, J. M. Comparison of two indicators of perceived egocentric distance under full-cue and reduced-cue conditions. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 72–85 (1997).
- Ooi, T. L., Wu, B. & He, Z. J. Distance determined by the angular declination below the horizon. *Nature* **414**, 197–200 (2001).
- Thomson, J. A. Is continuous visual monitoring necessary in visually guided locomotion? *J. Exp. Psychol. Hum. Percept. Perform.* **9**, 427–443 (1983).
- Rieser, J. J., Ashmead, D., Talor, C. & Youngquist, G. Visual perception and the guidance of locomotion without vision to previously seen targets. *Perception* **19**, 675–689 (1990).
- Loomis, J., DaSilva, J., Fujita, N. & Fukusima, S. Visual space perception and visually directed action. *J. Exp. Psychol.* **18**, 906–921 (1992).
- Loomis, J., DaSilva, J., Philbeck, J. W. & Fukusima, S. Visual perception of location and distance. *Curr. Dir. Psychol. Sci.* **5**, 72–77 (1996).
- Sinai, M. J., Ooi, T. L. & He, Z. J. Terrain influences the accurate judgement of distance. *Nature* **395**, 497–500 (1998).

- Loomis, J. M. & Knapp, J. M. in *Virtual and Adaptive Environments* (eds Hettinger, L. J. & Hass, M. W.) 21–46 (Erlbaum, Hillsdale, New Jersey, 2003).
- Hagen, M. A., Jones, R. K. & Reed, E. S. On a neglected variable in theories of pictorial perception: Truncation of the visual field. *Percept. Psychophys.* **23**, 326–330 (1978).
- Dolezal, H. *Living in a World Transformed: Perceptual and Performatory Adaptation to Visual Distortion* (Academic, New York, 1982).
- Loomis, J. M. & Philbeck, J. W. Is the anisotropy of perceived 3-D shape invariant across scale? *Percept. Psychophys.* **61**, 397–402 (1999).
- Loomis, J. M., Philbeck, J. W. & Zahorik, P. Dissociation of location and shape in visual space. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 1202–1212 (2002).
- Gibson, J. J. *The Perception of the Visual World* (Houghton, Mifflin, Boston, Massachusetts, 1950).
- Gibson, J. J. The perception of visual surfaces. *Am. J. Psychol.* **63**, 367–384 (1950).
- Freeman, R. B. Jr Effect of size on visual slant. *Psychol. Rev.* **72**, 501–504 (1956).
- Braunstein, M. L. Motion and texture as sources of slant information. *J. Exp. Psychol.* **78**, 247–253 (1968).
- Gibson, J. J. & Cornsweet, J. The perceived slant of visual surfaces—optical and geographical. *J. Exp. Psychol.* **44**, 11–15 (1952).
- Sedgwick, H. A. in *Handbook of Perception and Human Performance* (eds Boff, K. R., Kaufman, L. & Thomas, J. P.) 21.1–21.57 (Wiley, New York, 1986).
- Meng, J. C. & Sedgwick, H. A. Distance perception mediated through nested contact relations among surface. *Percept. Psychophys.* **63**, 1–15 (2001).
- Land, M. F. Scanning eye movements in a heteropod mollusc. *J. Exp. Biol.* **96**, 427–430 (1982).
- Land, M. F. Motion and vision: why animals move their eyes. *J. Comp. Physiol. A* **185**, 341–352 (1999).

Acknowledgements This research was supported in part by a grant from NIH to Z.J.H. and T.L.O. and by a RIG grant from the University of Louisville to Z.J.H.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Z.J.H. (zjhe@louisville.edu) or T.L.O. (tlooi@pcu.edu).

Inactivation of *hCDC4* can cause chromosomal instability

Harith Rajagopalan¹, Prasad V. Jallepalli^{1,2}, Carlo Rago¹, Victor E. Velculescu¹, Kenneth W. Kinzler¹, Bert Vogelstein^{1,3} & Christoph Lengauer¹

¹The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA

²Molecular Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

³Howard Hughes Medical Institute, The Johns Hopkins University Medical Institutions, Baltimore, Maryland 21231, USA

Aneuploidy, an abnormal chromosome number, has been recognized as a hallmark of human cancer for nearly a century¹; however, the mechanisms responsible for this abnormality have remained elusive. Here we report the identification of mutations in *hCDC4* (also known as *Fbw7* or *Archipelago*) in both human colorectal cancers and their precursor lesions. We show that genetic inactivation of *hCDC4*, by means of targeted disruption of the gene in karyotypically stable colorectal cancer cells, results in a striking phenotype associated with micronuclei and chromosomal instability. This phenotype can be traced to a defect in the execution of metaphase and subsequent transmission of chromosomes, and is dependent on cyclin E—a protein that is regulated by *hCDC4* (refs 2–4). Our data suggest that chromosomal instability is caused by specific genetic alterations in a large fraction of human cancers and can occur before malignant conversion.

CDC4 is an evolutionarily conserved E3 ubiquitin ligase that is thought to be involved in regulating the G1–S cell-cycle checkpoint by targeting proteins for destruction by the SCF complex of proteins². Mutations in the *hCDC4* gene were originally identified in a few breast and ovarian cancer cell lines and have been subsequently found in endometrial cancers^{5–7}. Here we have

addressed the role that *hCDC4* may have in colorectal tumorigenesis through mutational analysis and genetic dissection.

To determine whether *hCDC4* is genetically altered, its coding exons were amplified by polymerase chain reaction (PCR) from 190 colorectal tumours and sequenced. Somatic mutations in the *hCDC4* gene were found in 22 of the 190 samples. Nonsense mutations at codons 224, 278, 303, 393 and 446 were found in eight tumours. All of these mutations are predicted to result in truncation of the protein at a position amino-terminal to the fourth WD40 domain, thereby interrupting the binding of hCDC4 to its substrate (Fig. 1a). Six of the tumours had biallelic inactivations, either because of allelic loss of the wild-type gene (in three tumours) or because of two independent point mutations on opposite alleles (Supplementary Table 1). The 22 cancers with *hCDC4* mutations represent all stages of colorectal cancer. To determine the point in tumorigenesis at which the mutations occur, we sequenced *hCDC4* in adenomas—benign tumours that are the precursors of cancers. Out of 58 colorectal adenomas tested, four mutations were identified (Fig. 1a and Supplementary Table 1). Notably, two of these mutations occurred in adenomas less than 1 cm in diameter; such lesions have been shown to progress to malignancy only after 10–20 yr.

If the numerous missense changes in *hCDC4* inactivate the protein product, one would expect the mutations to cluster in key functional domains of the hCDC4 protein. Crystal structure coordinates of yeast CDC4 bound to a phosphopeptide of human cyclin E were used to reconstruct the spatial relationship between the two proteins^{8,9}. All but one of the missense mutations occurred in the WD40 domain repeats in the 3' region of the gene, and many resulted in the elimination of specific arginines that are thought to be involved in substrate-binding specificity^{8,10}. Notably, 18 of the 19 missense changes occurred at the surface of the eight-bladed β -propeller structure produced by the WD40 repeats (Fig. 1c, d). The lone remaining missense mutation occurred in the region encoding the F-box and may disrupt the binding of hCDC4 to the remainder of the SCF complex. Co-immunoprecipitation experiments also showed that cyclin E bound better to wild-type hCDC4 than to a mutationally altered protein (Fig. 1b), thus corroborating the suggestions from the structural data.

The several mutations identified above provide compelling evidence that *hCDC4* is involved in colorectal tumorigenesis. But how?

Some studies have suggested that hCDC4 controls cell proliferation by regulating replication, whereas others have proposed that hCDC4 may have a role in genetic instability^{6,11–13}. To determine the consequences of *hCDC4* inactivation in colorectal tumorigenesis, we used homologous recombination to disrupt both alleles of *hCDC4* in HCT116, a karyotypically stable colorectal carcinoma cell line (ref. 14 and Fig. 2a). Successful targeting was confirmed by PCR-based analysis of DNA and RNA (Fig. 2b, c). Two clones were obtained, each of which behaved identically in the assays described below.

The absence of *hCDC4* resulted in a high increase in cyclin E (Fig. 2d), indicating functional conservation of the hCDC4/cyclin E pathway originally identified in yeast and insect cells^{6,7}. The growth of *hCDC4*^{-/-} cells was indistinguishable from that of parental cells, and no obvious differences were observed in plating efficiency, cell-cycle profile or saturation density (data not shown); however, microscopic examination revealed several marked cytological abnormalities. First, *hCDC4*-deficient cells showed nuclear atypia characterized by micronuclei and lobulated or elongated nuclei, none of which was observed in parental wild-type cells (18.3% versus 1.0%, $P < 10^{-5}$, $n = 400$ cells for each cell type; Fig. 3a–d). The fraction of cells with such aberrant nuclei did not vary with the generational age of the cells, as such cells were found in both early and late passage cultures (Supplementary Table 2 and data not shown).

Second, immunofluorescence staining showed an increase in the frequency of multipolar spindles in *hCDC4*^{-/-} cells (Fig. 3e, f), a mitotic defect that was rarely observed in wild-type cells (7% versus 0.5%; $P < 0.001$). Fluorescence *in situ* hybridization (FISH) of a pan-centromeric probe to interphase nuclei showed that a few micronuclei had one or two signals, but most micronucleated chromosomal material lacked detectable centromeric DNA (Fig. 3c). Notably, none of the *hCDC4*^{-/-} knockout clones showed an increase in apoptosis (data not shown), indicating that the generation of micronuclei was not likely to be fatal in cellular replication.

To elucidate the basis of these cellular abnormalities, we used time-lapse videomicroscopy of *hCDC4*^{-/-} and parental cells to visualize chromosomes in living cells¹⁵. Eight of twenty-six (31%) mitoses in *hCDC4*^{-/-} cells did not execute chromosome congression during metaphase ($P < 0.002$; Supplementary Fig. 1 and

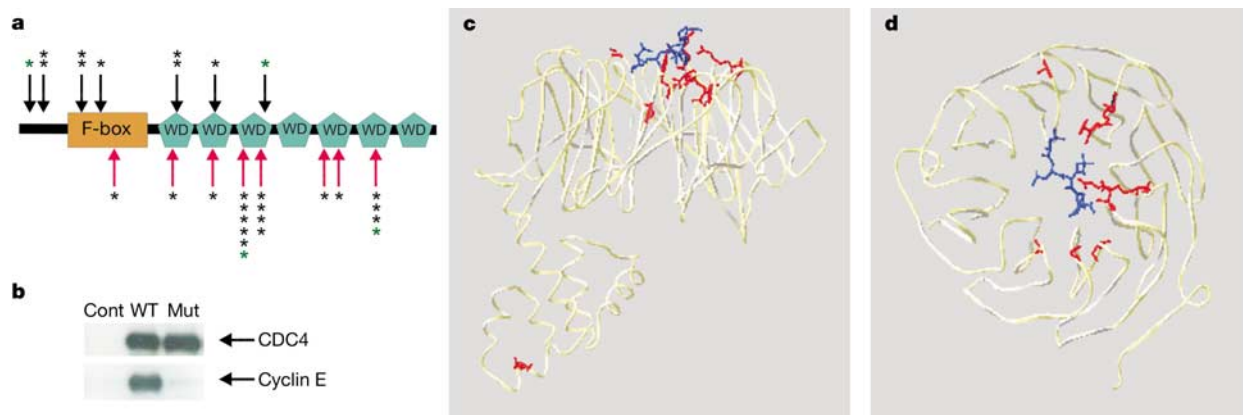


Figure 1 *hCDC4* mutations in colorectal cancers and adenomas. **a**, Location (arrows) and frequency (asterisks) of mutations in *hCDC4*. Black arrows indicate nonsense mutations, red arrows indicate missense mutations, black asterisks indicate the number of mutations found in cancers, green asterisks show mutations found in adenomas. **b**, Western blotting for Myc (Myc-tagged CDC4) and cyclin E after immunoprecipitation with an antibody against Myc. Cont, untransfected cells; WT, cells co-transfected with cyclin E and

Myc-tagged wild-type CDC4; Mut, cells co-transfected with cyclin E and Myc-tagged mutant CDC4 (Arg465Cys). **c**, **d**, Three-dimensional views of yeast CDC4 bound to human cyclin E^{8,9}. CDC4 is shown in yellow, cyclin E in blue. Amino acids in CDC4 that align with residues affected by missense mutations in hCDC4 are labelled in red. **c**, Longitudinal view. **d**, View down the z-axis.

Supplementary Movie). The full chromosome complement did not align on the metaphase plate despite numerous 'attempts', and mitosis ultimately proceeded with an unequal segregation of DNA content. In two cells, the formation of a micronucleus was observed during filming. The duration of mitosis was extremely prolonged in all cells, lasting 147 ± 9 min. The remaining 70% of mitoses in $hCDC4^{-/-}$ cells lasted 58 ± 2.5 min, similar to those observed in parental cells, and were executed in a morphologically normal manner. No other obvious differences in cell-cycle execution were observed.

To determine whether this marked phenotype was reproducible, we disrupted the $hCDC4$ gene in DLD-1, another karyotypically stable colorectal cancer cell line. We obtained a clone in which both alleles were disrupted (Fig. 2b, c), and the phenotype of this clone

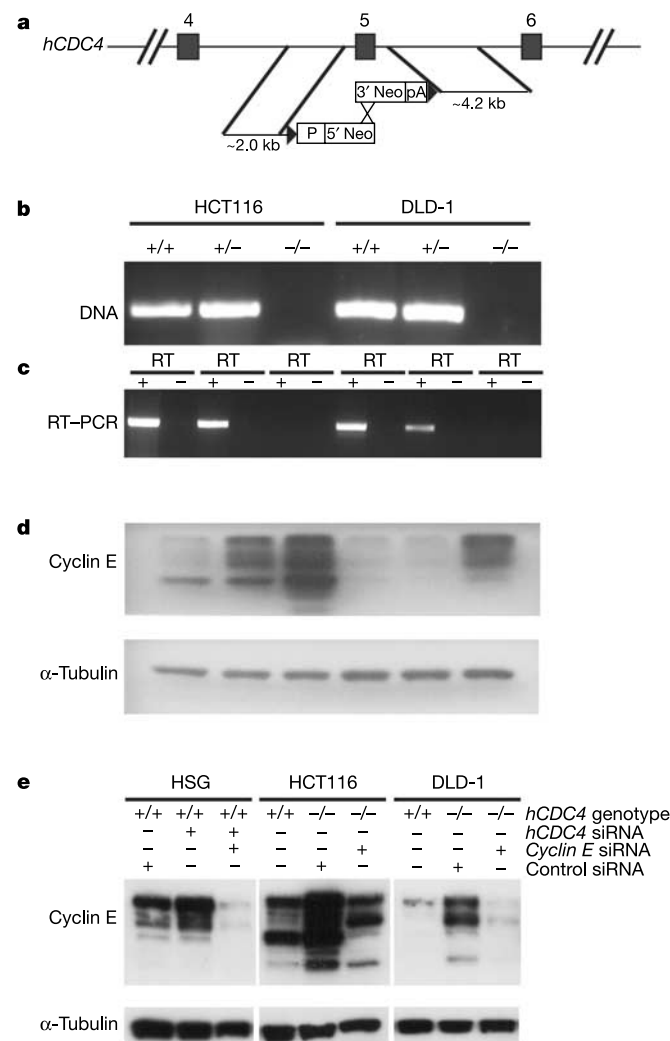


Figure 2 Disruption of $hCDC4$ and cyclin E. **a–d**, Knockout of $hCDC4$ by homologous recombination. **a**, Exon 5 was targeted for deletion by homologous recombination using 5' and 3' arms in a bipartite vector system. **b**, Confirmation of knockouts by a PCR screen of genomic DNA. **c**, Reverse-transcription PCR screen of cDNA. **d**, Western blotting with an antibody against cyclin E. **e**, siRNA targeting of $hCDC4$ and/or cyclin E. The effects of targeting were assessed by western blotting with an antibody against cyclin E. Control siRNA indicates mock transfection in HSG and HCT116 $hCDC4^{-/-}$ cells and transfection with Lamin A/C siRNA in DLD-1 $hCDC4^{-/-}$ cells.

was indistinguishable from that of HCT116 $hCDC4^{-/-}$ cells. In particular, there were no obvious growth defects in $hCDC4^{-/-}$ cells, but there was a significant increase in cyclin E protein (Fig. 2d) and a higher percentage of micronuclei formation as compared with parental cells (22.0% versus 0.9%, $P < 10^{-5}$, $n > 400$ cells for each; Supplementary Table 2).

We examined whether the $hCDC4^{-/-}$ cells had acquired chromosomal instability by carrying out multicolour FISH with centromeric probes for chromosomes 7, 17, 18 and X on interphase nuclei once cells had gone through 25 generations after clone isolation. Deviations from the normal number of chromosomes were frequently observed in $hCDC4^{-/-}$ cells derived from either HCT116 or DLD-1, but not in their $hCDC4^{+/+}$ heterozygous or wild-type parental counterparts (Table 1). The degree of instability in the $hCDC4^{-/-}$ cells is similar to that observed in chromosomally unstable colorectal cancer cells¹⁶. Notably, the abnormal chromosome numbers were most commonly observed in cells that possessed micronuclei or that showed other abnormal nuclear morphology (Fig. 3a, b, d).

It was possible that the chromosomal instability phenotype observed in the two $hCDC4^{-/-}$ cancer lines was caused by mutations in other genes in the parental lines used to generate the knockout lines. To investigate the effects of $hCDC4$ inactivation alone, we used transient transfection of short interfering RNA (siRNA) duplexes to suppress specifically the expression of $hCDC4$

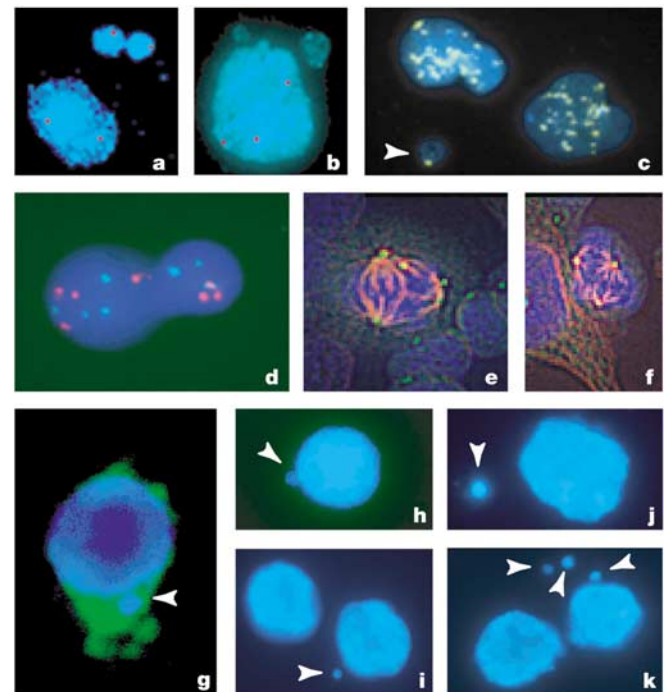


Figure 3 $hCDC4$ -deficient human cells are chromosomally unstable. **a, b**, $hCDC4^{-/-}$ HCT116 cells hybridized with a chromosome-17-centromere-specific FISH probe (red dots) and counterstained for DNA content with DAPI. **c**, $hCDC4^{-/-}$ cells hybridized with a pan-centromeric FISH probe (green), showing centromeric DNA in a micronucleus (arrowhead). **d**, $hCDC4^{-/-}$ DLD-1 cells hybridized with FISH probes specific for centromeres of chromosome 7 (red) and 18 (green). **e, f**, Spindle dysfunction in $hCDC4^{-/-}$ HCT116 cells. Cells were stained with antibodies to α -tubulin (red) and γ -tubulin (green) to detect mitotic spindles and centrosomes, respectively. **g**, Cyclin E immunofluorescence (green) and DAPI counterstaining (blue) of a DLD-1 cell expressing exogenous cyclin E, showing a micronucleus (arrowhead). **h–k**, Micronuclei (arrowheads) are found in HSG cells either transfected with siRNA targeted towards $hCDC4$ (**h, i**), or overexpressing exogenous cyclin E (**j, k**).

in cells derived from normal human salivary gland epithelium (HSG). A 40% decrease in *hCDC4* messenger RNA (data not shown) was sufficient to produce an increased abundance of cyclin E protein (Fig. 2e). These *hCDC4* 'knockdown' cells also showed a significant increase in the percentage of micronucleus formation (9.0% versus 3.8%, $P < 0.0001$, $n = 1,000$ cells for each; Fig. 3h, i, and Supplementary Table 2). It should be noted that the assessment of micronuclei as a proxy for chromosomal instability suffers from some theoretical disadvantages, but these are more problematic in long-term experiments than in the experiments described here¹⁷.

Cyclin E is the only known target of the ubiquitin ligase activity of *hCDC4*. To determine whether the increased abundance of cyclin E was sufficient to reproduce the phenotype observed on inactivation of *hCDC4*, we overexpressed cyclin E in DLD-1 and HSG cells through transient transfection. An increase in the percentage of micronuclei formation (Fig. 3j, k) was observed in both types of cells as compared with cells transfected with a control expression vector (9.3% versus 3.0% for DLD-1 and 9.7% versus 3.6% for HSG, $P < 0.0001$, $n > 400$ cells for each; Supplementary Table 2). We also carried out immunofluorescence in DLD-1 cells transfected with a cyclin E expression vector, and counted micronuclei in cells staining brightly for cyclin E (Fig. 3g). We found that cells expressing large quantities of cyclin E produced significantly more micronuclei than did cells without cyclin E staining (21% versus 6%, $P < 0.001$, $n > 100$ cells for each; Supplementary Table 2).

Having shown that overexpression of cyclin E could recapitulate the micronucleus phenotype observed in cells with inactivated or suppressed *hCDC4*, we next determined whether cyclin E was necessary to produce this phenotype. Knockdown of *cyclin E* expression in the cancer lines in which *hCDC4* was knocked out significantly reduced the percentage of micronuclei formation (7.9% versus 17.5% for HCT116, and 8.5% versus 15.2% for DLD-1, $P < 0.0001$, $n = 1,000$ cells for each; Fig. 2e and Supplementary Table 2). Similar results were observed in HSG cells: when co-transfected with siRNA duplexes targeting both *hCDC4* and *cyclin E*, the percentage of micronucleus formation was much less than that observed when only *hCDC4* expression was knocked down (2.7% versus 9%, $P < 0.0001$, $n = 1,000$ cells for each; Supplementary Table 2). Taken together, these results imply that an upregulation of cyclin E is essential for the chromosomal instability phenotype observed after inactivation of *hCDC4*.

Table 1 Measurement of chromosomal instability by FISH

Cell line	Chromosome	Signals (%)					
		0	1	2	3	4	5
HCT116 parental	7		1	99			
	17		2	98			
	X		99	1			
HCT116 <i>hCDC4</i> ^{+/-}	7		1.5	97	1	0.5	
	17		1.5	98		0.5	
	X	0.5	98.5	1			
HCT116 <i>hCDC4</i> ^{-/-}	7		5.5	88	5	1	0.5
	17	1	7	88	2	2	
	X	1.5	89	9.5			
DLD-1 parental	7			98		2	
	17		2	97		1	
	18		3	93	2	2	
DLD-1 <i>hCDC4</i> ^{+/-}	X		98	2			
	7			99		1	
	17		4	95		1	
DLD-1 <i>hCDC4</i> ^{-/-}	18		3	95	1	1	
	X		98	2			
	7	2	4	88	4	2	
	17		14	83	2	1	
	18	2	3	89	4	2	
	X	5	86	5		4	

A total of 200 nuclei were examined for FISH signals from each of the indicated centromeric probes. $P < 10^{-7}$ for the percentage of signals off the mode in *hCDC4*^{-/-} cells as compared with HCT116 wild-type or DLD-1 wild-type (*hCDC4*^{+/+}) cells.

Disparate cellular events and pathways have been proposed to have a role in the genetic instability found in cancer cells¹⁸. Only a few cancers have been shown to contain specific mutations that can promote aneuploidy¹⁹, and the stage at which chromosomal instability occurs in tumorigenesis remains controversial²⁰⁻²⁶. Our results suggest that mutational inactivation of *hCDC4* is likely to be a chief cause of chromosomal instability in cancers and that this mutation can occur at an early stage. These data lend support to the hypothesis that the chromosomal instability of cancer cells, like their abnormal growth control, is a direct result of specific genetic alterations. □

Methods

Sequencing

Primers for PCR amplification and direct sequencing of exons 2–10 of *hCDC4* were designed with the NetPrimer program (Premier Biosoft) and synthesized by GeneLink. PCR amplification and sequencing were done on tumour DNA from 58 adenomas and 190 early passage colorectal cancer cell lines passaged as xenografts in nude mice or *in vitro* as described²⁷. Sequencing was done by using 384-capillary automated sequencing apparatus (SpectruMedix). We confirmed that all mutations were somatic by amplifying and sequencing patient-matched normal DNA. Sequence traces were analysed with the Mutation Explorer software package (SoftGenetics).

Knockout and siRNA

General methods for gene disruption in colorectal cancer cells using a bipartite vector system have been described¹⁴. Constructs and primer sequences are available from the authors on request.

The coding strand of the RNA oligonucleotide duplex directed to *hCDC4* was 5'-GAGUUGGCACUCUAUGUGC-3' and had 3' dTdT overhangs (Dharmacon). We purchased siRNA duplexes for Lamin A/C from Dharmacon. siRNA duplexes targeting cyclin E were constructed by using a Shortcut RNAi kit (NEB) in accordance with the manufacturer's protocols. Primer sequences are available from the authors on request. We obtained HSG cells²⁸ from A. Rosen and grew them in DMEM medium supplemented with 10% fetal bovine serum. RNA duplexes were transfected into cells by using Oligofectamine (Invitrogen) in accordance with the manufacturer's instructions. Cellular protein, mRNA and nuclear structure were assayed 72 h after transfection. We determined mRNA levels by real-time PCR and evaluated nuclear structure after staining with 4',6-diamidino-2-phenylindole dihydrochloride (DAPI).

Protein expression and immunoprecipitation

For overexpression of cyclin E, the cyclin E open reading frame was cloned into pBI-EGFP¹⁹ and transiently transfected into DLD-1 cells (American Type Culture Collection) by using Lipofectamine (Invitrogen). For immunoprecipitation, a fragment of *hCDC4* complementary DNA, nucleotides 374 to 1,874, was amplified from DLD-1 cDNA and cloned in-frame into the pCMV-Myc (Clontech) multiple-cloning site using the *Bgl*III and *Not*I restriction sites. The Arg465Cys mutation was introduced by sexual PCR and cloned into the vector in a similar way. Primer sequences are available from the authors on request.

Cyclin E full-length cDNA was cloned into pIRES-Hyg2 (Clontech). Cloned constructs were co-transfected into DLD-1 cells with Lipofectamine (Invitrogen) at a 1:3 or 1:10 (mass:mass) ratio of Myc-*hCDC4*:cyclin E. Crude extracts were collected after 24 h by lysing cells in ice-cold buffer containing 50 mM Tris, 0.5% deoxycholate, 1% Nonidet P-40, 150 mM NaCl and Complete protease inhibitors (Roche) and subsequently incubated with agarose beads conjugated to mouse monoclonal antibody directed towards the c-Myc epitope tag (9E10, Santa Cruz Biotech).

Immunoblotting

Protein extracts from HCT116, DLD-1 and derivative lines were prepared by lysing cells on ice in RIPA buffer (Boston BioProducts). We briefly sonicated the extracts and removed debris by centrifugation at 10,000g for 15 min at 4 °C. Immunoblotting was done by using Immobilon P membranes (Millipore) according to the manufacturer's instructions. Commercially available antibodies to cyclin E (Santa Cruz Biotech) and c-Myc (Roche) were used as recommended by the manufacturer. Signals were developed using the ECL Advance western blot detection kit (Amersham).

Immunofluorescence

For α -tubulin and γ -tubulin staining, cells grown on chamber slides were fixed in 4% paraformaldehyde for 10 min. After permeabilization in 0.1% Triton X-100, cells were blocked in a solution of 4% goat serum in PBS, and then incubated with monoclonal antibodies to α -tubulin (IgG2a) and γ -tubulin (IgG1), followed by isotype-specific antibodies coupled to Alexa 594 and Alexa 488 (Molecular Probes). Images were acquired and deconvoluted by the MetaMorph software package (Molecular Devices). For cyclin E staining, cells grown on slides were fixed in ice-cold methanol for 10 min. Cells were blocked with 10% goat serum in PBS, and incubated first with monoclonal antibodies to cyclin E (HE111, Santa Cruz Biotech) and then with a fluorescein isothiocyanate (FITC)-conjugated secondary antibody against mouse IgM and IgG (Pierce). Fixed slides were stained with DAPI by using traditional methods.

Time-lapse imaging

Cells were stably transfected with pBOS-histone-H2B-GFP¹⁵ and analysed on a TE200 inverted microscope (Nikon) with a ×40 objective enclosed in a temperature- and carbon-dioxide-controlled incubator (Neve). Images were acquired with a CCD camera (Princeton) at intervals ranging from 1 to 10 min and analysed by the MetaMorph software program (Molecular Devices).

Fluorescent *in situ* hybridization

Methods for FISH analysis with chromosome-specific centromeric probes and quantitative analysis of chromosome loss rates have been described¹⁶. A pan-centromeric FISH probe (IDBright Pan-Centromeric, ID Labs) was used in accordance with the manufacturer's instructions. Multicolour FISH with centromeric probes for chromosomes 7, 17, 18 and X was also done on interphase nuclei once cells had gone through 25 generations after clone isolation. To prepare metaphase spreads, we treated cells with 0.1 µg ml⁻¹ colcemid (KaryoMax, Invitrogen) for 1 h and processed them by standard methods. Karyotyping was done by standard G-banding procedures.

Received 17 September 2003; accepted 5 January 2004; doi:10.1038/nature02313.

1. Boveri, T. *Zur Frage der Entstehung maligner Tumoren* (Gustav Fischer, Jena, Germany, 1914).
2. Winston, J. T., Koepf, D. M., Zhu, C., Elledge, S. J. & Harper, J. W. A family of mammalian F-box proteins. *Curr. Biol.* **9**, 1180–1182 (1999).
3. Schwab, M. & Tyers, M. Cell cycle. Archipelago of destruction. *Nature* **413**, 268–269 (2001).
4. Bartek, J. & Lukas, J. Cell cycle. Order from destruction. *Science* **294**, 66–67 (2001).
5. Moberg, K. H., Bell, D. W., Wahrer, D. C., Haber, D. A. & Hariharan, I. K. Archipelago regulates Cyclin E levels in *Drosophila* and is mutated in human cancer cell lines. *Nature* **413**, 311–316 (2001).
6. Spruck, C. H. *et al.* hCDC4 gene mutations in endometrial cancer. *Cancer Res.* **62**, 4535–4539 (2002).
7. Strohmaier, H. *et al.* Human F-box protein hCdc4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line. *Nature* **413**, 316–322 (2001).
8. Orlicky, S., Tang, X., Willems, A., Tyers, M. & Sicheri, F. Structural basis for phosphodependent substrate selection and orientation by the SCF^{Cdc4} ubiquitin ligase. *Cell* **112**, 243–256 (2003).
9. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
10. Koepf, D. M. *et al.* Phosphorylation-dependent ubiquitination of cyclin E by the SCF^{Fbw7} ubiquitin ligase. *Science* **294**, 173–177 (2001).
11. Spruck, C. H., Won, K. A. & Reed, S. I. Deregulated cyclin E induces chromosome instability. *Nature* **401**, 297–300 (1999).
12. Nash, P. *et al.* Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* **414**, 514–521 (2001).
13. Goh, P. Y. & Surana, U. Cdc4, a protein required for the onset of S phase, serves an essential function during G₂/M transition in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19**, 5512–5522 (1999).
14. Jalileppä, P. V. *et al.* Securin is required for chromosomal stability in human cells. *Cell* **105**, 445–457 (2001).
15. Kanda, T., Sullivan, K. F. & Wahl, G. M. Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells. *Curr. Biol.* **8**, 377–385 (1998).
16. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386**, 623–627 (1997).
17. Savage, J. R. Acentric chromosomal fragments and micronuclei: the time-displacement factor. *Mutat. Res.* **225**, 171–173 (1989).
18. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
19. Cahill, D. P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392**, 300–303 (1998).
20. Thiagalingam, S. Evaluation of chromosome 18q in colorectal cancers. *Nature Genet.* **13**, 343–346 (1996).
21. Pihan, G. A. *et al.* Centrosome defects and genetic instability in malignant tumors. *Cancer Res.* **58**, 3974–3985 (1998).
22. Rajagopalan, H., Nowak, M. A., Vogelstein, B. & Lengauer, C. The significance of unstable chromosomes in colorectal cancer. *Nature Rev. Cancer* **3**, 695–701 (2003).
23. Sieber, O. M., Heinemann, K. & Tomlinson, I. P. Genomic instability—the engine of tumorigenesis? *Nature Rev. Cancer* **3**, 701–708 (2003).
24. Nowak, M. A. *et al.* The role of chromosomal instability in tumor initiation. *Proc. Natl Acad. Sci. USA* **99**, 16226–16231 (2002).
25. Sieber, O. M. *et al.* Analysis of chromosomal instability in human colorectal adenomas with two mutational hits at APC. *Proc. Natl Acad. Sci. USA* **99**, 16910–16915 (2002).
26. Shih, I. M. *et al.* Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. *Cancer Res.* **61**, 818–822 (2001).
27. Bardelli, A. *et al.* Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300**, 949 (2003).
28. Shirasuna, K., Sato, M. & Miyazaki, T. A neoplastic epithelial duct cell line established from an irradiated human salivary gland. *Cancer* **48**, 745–752 (1981).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Meszler and L. Morsberger for technical assistance; A. Rosen and S. White for reagents; and the members of our Center for Cancer Genetics and Therapeutics for help and support. This work was supported by the Clayton Fund, NIH grants, and a Translational Research Award from the Virginia and D. K. Ludwig Fund for Cancer Research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.L. (lengauer@jhmi.edu).

Spatially restricted microRNA directs leaf polarity through ARGONAUTE1

Catherine A. Kidner & Robert A. Martienssen

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

Gene regulation by RNA interference requires the functions of the PAZ domain protein Argonaute. In plants, mutations in *ARGONAUTE1* (*AGO1*) are associated with distinctive developmental defects that suggest a role for microRNA (miRNA) in organ polarity. Potential targets of miRNA regulation are the homeodomain/leucine zipper genes *PHABULOSA* (*PHB*) and *PHAVOLUTA* (*PHV*)¹. These genes are expressed in a polar fashion in leaf primordia and are required for adaxial cell fate^{2,3}. Here we show that a 21-nucleotide miRNA that directs cleavage of *PHB/PHV* messenger RNA accumulates first in the embryonic meristem, and then in the abaxial domain of the developing leaf. miRNA distribution is disrupted by mutations in *AGO1*, indicating that *AGO1* affects the regulation of miRNA. In addition, interactions between homeodomain/leucine zipper genes and an allelic series of *ago1* indicate that miRNA acts as a signal to specify leaf polarity.

Surgical isolation of early leaf primordia results in outgrowth of 'centric' or radial organs, suggesting that a meristem-derived signal directs adjacent regions of the leaf primordia to adopt an adaxial fate, and directs the regions of the leaf primordia furthest from the meristem to become abaxial⁴. In *Arabidopsis*, mutants in *AGO1* (ref. 5) and the related gene *PINHEAD/ZWILLE* (*PNH/ZLL*)^{6,7} have overlapping defects in meristem maintenance and leaf polarity consistent with a defect in this signalling process, and the two genes are partially redundant⁶. However, whereas *ago1* leaves are adaxialized (see below), *pnh* leaves are weakly abaxialized⁶.

Argonaute genes are defined by carboxy-terminal PAZ and PIWI domains, and are widespread in eukaryotes⁸. In *Drosophila*, mutations in *sting/auvergne* and *dAGO1* have defects in germline development⁹, and *Sting/Auvergne* is required for translation of *oskar* in the polar granules¹⁰. *Argonaute* genes are required for post-transcriptional gene silencing and RNA interference (RNAi)^{8,11}, and *AGO1*-like proteins are part of the RISC (RNA interference silencing complex), which targets mRNA for degradation using miRNA as a guide^{12,13}. miRNAs from *Arabidopsis* match several transcription factor genes, potentially accounting for *ago1* and *pnh/zll* phenotypes^{1,14}.

We isolated an allelic series of *AGO1* to investigate its role in polarity (Fig. 1). The strong alleles *ago1-9* and *ago1-10* are frame-shift truncations of 384 and 404 amino acids respectively, deleting the PAZ and PIWI domains. Vegetative organs are strap-like or radial^{5,6}, with radial organs (*ago1-10*) or reduced flowers (*ago1-9*) in the inflorescence (Fig. 1), suggesting a complete loss of adaxial/abaxial polarity. Ectopic ovules are found occasionally (*ago1-9*) or frequently (*ago1-11* and *ago1-12*) on the outside (abaxial) carpel wall, usually near the base (Fig. 1c). *ago1-11* and *ago1-12* are both weak alleles that retain the PAZ domain. *ago1-11* has an A-to-T conversion at the splice acceptor site of the fourteenth intron, leading to mis-splicing and exon skipping. The first pair of leaves are sometimes radial or trumpet-shaped but the remainder are flattened. Trichomes appear on the abaxial rather than the adaxial side of early leaves, suggesting an inversion of epidermal polarity in this weak allele (Fig. 1d). *ago1-12* has a C-to-T transition at nucleotide 2,414, resulting in a leucine substitution for an absolutely conserved histidine residue in the PIWI domain. *ago1-12* leaves are almost always trumpet-shaped with glossy dark-green tissue (adaxial) outside and reflective epidermis (abaxial) inside the

bell (Fig. 1b). This phenotype resembles the dominant mutant *Phabulosa1-D*³.

Enhancer-trap reporter ET2689 is inserted in an intergenic region upstream of *At3g28925* and is expressed on the proximal abaxial side of leaves. In *ago1-10*, expression is lost from the leaves and appears ectopically in the apical hook (Fig. 2a, b). ET3964 is expressed on the adaxial side of leaves and flowers, as well as in axils, and is inserted in the 5' UTR of *At1g67260*, a TCP gene¹⁵. Other members of the TCP gene family are regulated by the JAW miRNA and control growth¹⁶. In *ago1-10*, ET3964 is ectopically expressed and upregulated (Fig. 2c, d), suggesting that *At1g67260* is also regulated by RNAi. *FILAMENTOUSFLOWER* (*FIL*) and *KANADII* (*KAN1*) are expressed abaxially in wild-type organs and regulate polarity^{17–20}.



Figure 1 Mutations in *AGO1* are adaxialized and mutations in *REV* are enhanced by *ago1* and *dcl1-9*. **a–d**, Six-week-old plants. **a**, *ago1-10* (*ago1-9* has a very similar phenotype); **b**, *ago1-12*; **c**, **d**, *ago1-11*. **c**, In wild-type carpels, ovules are formed only on the adaxial surface; however, in *ago1-11*, ovules emerge from the abaxial surface as well. **d**, First or second leaves of *ago1-11*. At this stage, wild-type leaves have trichomes on the adaxial side only, and *ago1-11* has trichomes abaxially. **e**, *ago1-11/ago1-11, dcl1-9/dcl1-9* seedlings lack meristem function and organ polarity. **f**, *ago1-10/rev-6* double mutants lack a shoot apical meristem and cotyledons are reduced or absent. **g**, *dcl1-9/rev-6* mutants have narrow, epinastic leaves and form filamentous organs in the place of flowers. **h**, *Phb1-D/Phb1-D, ago1-10/ago1-10* seedlings lack meristem function and have no, or only rudimentary, organs. **i**, *Phb1-D/+ , ago1-12/+* seedlings have an enhanced polar defect, resembling *Phb1-D* homozygotes.

FIL is expressed at a much lower level in *ago1-9* but *KAN1* is expressed at normal levels (Fig. 2e).

The closely related homeodomain/leucine zipper (HD-ZIP III) genes *PHB* and *PHV* convey leaf polarity². *PHB* and *PHV* are located in segmental genome duplications²¹ and are probably redundant. Loss-of-function mutants in *PHB* and *PHV* have no phenotype²², although the closely related gene *REVOLUTA* (*REV*) is required for axillary meristem development and vascular patterning²³, and the *phb,phv,rev* triple mutant has severe loss of polarity and of meristem function²². *PHB* negatively regulates *FIL*, and both *REV* and *PHV* are negatively regulated by *KAN1* and *KAN2* (refs. 17,20). In early embryos *PHB* and *REV* are expressed in the shoot apical meristem (SAM), in adaxial domains of developing cotyledons, and in vasculature^{23,2}. *PHB* expression is lost from the SAM of torpedo-stage embryos but is regained late in embryogenesis, extending in 'rays' to young primordia after germination, and becoming progressively restricted to the adaxial domain and the vasculature². *REV* expression forms an inverted-cup pattern within the SAM. In organ primordia *REV* is restricted to the vasculature and to a broader adaxial domain than *PHB*. The combined expression pattern of both genes overlaps with that of *PNH/ZLL*. *AGO1* is expressed uniformly in embryos, SAMs and developing primordia⁶.

Two miRNAs from *Arabidopsis*, miR165 and miR166 (ref. 24), match *PHB*, *PHV* and *REV* mRNA, and direct their cleavage^{22,25}. The 19-base-pair (bp) miRNA match is specifically disrupted in each of

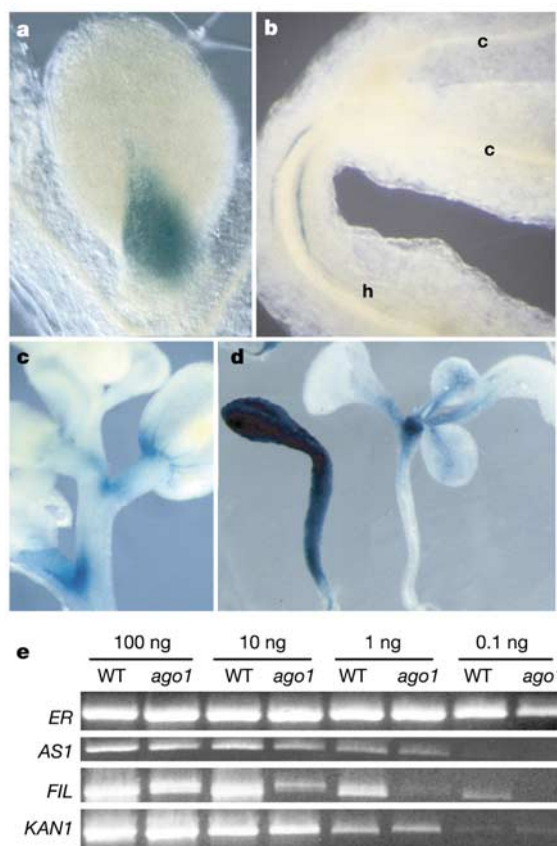


Figure 2 Reporter gene expression in *ago1-10* demonstrates adaxialization. **a**, ET2689 is expressed on the abaxial proximal side of the leaf. **b**, In *ago1-10*, no expression is seen in the leaves but two stripes of expression appear at the apical hook (c, cotyledon; h, hypocotyl). **c**, **d**, ET3964 is expressed on the adaxial side of the leaf and cotyledon axil (c and right side of d). In *ago1-10* it is overexpressed (left side of d). **e**, *FIL* is expressed at lower levels in *ago1-10*. A dilution series of total RNA from the wild type and *ago1-10* has consistently lower amplification by RT-PCR when using primers specific for *FIL* but not for *ERECTA* (*ER*), *ASYMMETRIC LEAVES1* (*AS1*) or *KANADII* (*KAN1*).

nine independent dominant mutations in *PHB* and *PHV* that have strongly adaxialized organs and form trumpet-shaped leaves². The dominant mutations in *PHB* lead to ectopic expression², and transcripts of the dominant mutants cannot be cleaved *in vitro*²⁵.

Both *ago1-12* and *Phb1-D* have trumpet-shaped leaves with adaxial tissue on the outside. These similarities indicated that the *ago1* phenotype might be due to failure to downregulate *PHB* and *PHV* mRNA. Consistent with this view, the *Phb1-D* gain-of-function phenotype was enhanced by *ago1*, presumably by upregulation of *PHV* due to loss of miRNA regulation (Fig. 1h, i). However, loss-of-function mutants in the *PHB* homologue *REV* were also strongly enhanced (Fig. 1f). In *ago1-10;rev-6* double mutants, SAMs and cotyledons fail to develop at the heart stage, although root growth is unaffected (Fig. 1f). We explored this further by examining *rev-6* in combination with *dcl1-9*. *DCL1* encodes a homologue of the double-stranded RNase DICER and fails to accumulate miRNA²⁴. *ago1-10;dcl1-9* double mutants are lethal at the embryo stage (data not shown) and *ago1-11;dcl1-9* double mutants lose meristem function and organ polarity (Fig. 1e). This confirms that the RNAi pathway is required for polarity. *rev-6;dcl1-9* double mutants had filamentous flowers resembling *rev;fil* (ref. 26), and *FIL* was severely downregulated in *rev-6;dcl1-9* inflorescences (Fig. 1g and Supplementary Information). This loss of abaxial fate would be expected from the loss of miRNAs regulating *PHB* and *PHV* but is only revealed in the absence of *REV*.

ago1 and *dcl1* have different effects on the accumulation of miRNA²⁷, perhaps accounting for their different interactions with *rev*. We examined this further by *in situ* hybridization using the miRNA precursor as a probe (Fig. 3). Consistent with northern blot analysis²⁷, hybridization was only detected with probes that included the miRNA (Fig. 3a and Supplementary Information). In bent cotyledon-stage wild-type embryos, the miRNA accumulates in the vasculature and in the SAM (Supplementary Information), but in germinating seedlings, the miRNA accumulates specifically in the abaxial domain of developing leaf primordia (Fig. 3a). Weaker hybridization can be detected in the SAM (Fig. 3a). This pattern thus limits miRNA-directed cleavage of *PHB/PHV/REV* to the abaxial side of young primordia, and is largely complementary to the expression patterns of these genes². miR165 was not detected in P0, although *Phb1-D* is ectopically expressed in incipient primordia. It is possible that low levels could not be detected in these cells.

To determine whether miRNA accumulation influenced leaf polarity through this mechanism, we examined *PHB* expression. In the wild type, *PHB* is expressed on the adaxial side of the leaf, as previously reported² (Fig. 3c). In *ago1-10*, *PHB* is ectopically expressed throughout leaf primordia, accounting for adaxialization of the leaf (Fig. 3d). However, miR165 is also ectopically expressed in the adaxial domain of *ago1*, and is lost from the meristem (Fig. 3c). The ectopic accumulation of miRNA in *ago1* may be due to failure to enter the RNAi pathway. Recently, the PAZ domain has been shown to bind miRNA and may be important in miRNA maturation and incorporation into the RISC²⁸.

The pattern of accumulation of miR165 accounts for the different roles of *PNH/ZLL* and *AGO1* in leaf polarity, namely that *ago1* is adaxialized whereas *pnh/zll* is weakly abaxialized. *PNH/ZLL* is expressed in the adaxial domain⁶, where it could not direct *PHB/PHV* cleavage because of the absence of miRNA (Fig. 3f). However, *PNH/ZLL* could direct cleavage in *ago1*, as miR165 is mis-localized to the adaxial domain. Reduction of adaxial *PHB* in *ago1* would enhance loss-of-function mutants in *REV*. Conversely, *pnh-2* would have no effect, and is indeed epistatic to *rev-6* (data not shown). *PNH/ZLL* and *AGO1* may even positively regulate *PHB/PHV/REV* in the absence of miRNA, by analogy with *aubergine* and *oskar* in *Drosophila*¹⁰. Thus differential accumulation of miRNA accounts for differences between *pnh* and *ago1*, and their interactions with *rev*.

Sussex⁴ proposed that a signal passed between the meristem and leaf primordia to convey organ polarity. Information also flows from the leaf to the meristem, as loss of organ polarity can result in meristem arrest^{2,3,17}. In tobacco, disruption of RNA trafficking using viral movement proteins leads to radial leaves that may be at least partially adaxialized, indicating a mobile RNA signal²⁹. Our results indicate that miR165 and miR166, or their precursors, contribute to this polarizing signal: miRNA initially accumulates in the meristem, becoming localized to the abaxial domain during leaf development. This process requires *AGO1*, indicating that miRNA localization is, at least in part, under post-transcriptional control. The loss of

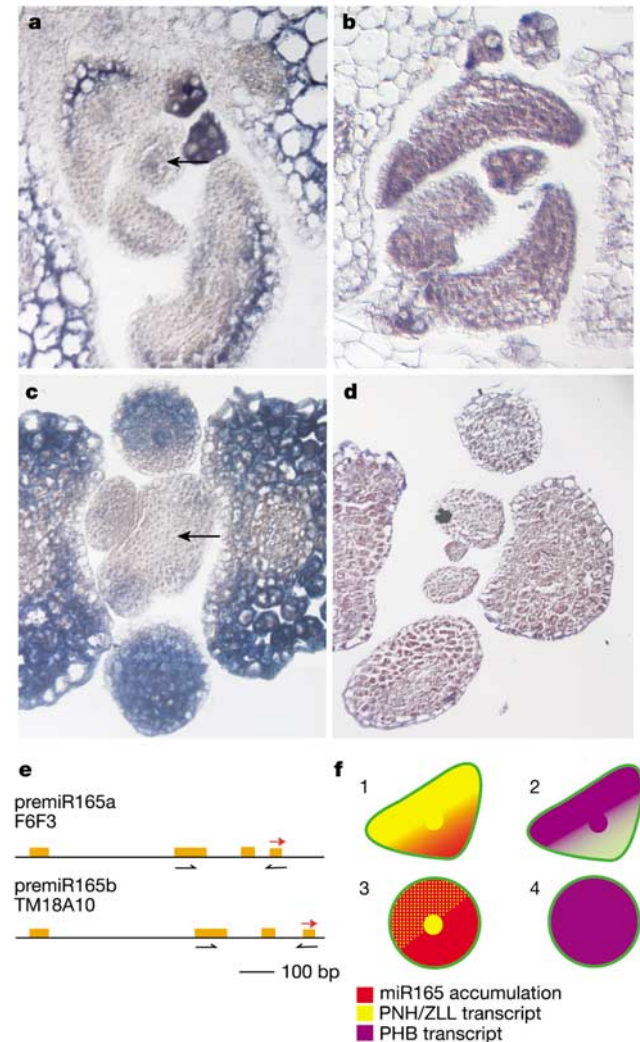


Figure 3 Transcripts from miR165 accumulate in the embryonic meristem and on the abaxial side of the leaf. **a**, In the wild type, miR165 accumulates on the abaxial side of 7-day-old leaves at stages P1, P2 and P3 and also weakly in the meristem (arrow). **b**, *PHB* accumulates on the adaxial side of the leaf primordia in the wild type. **c**, In *ago1-10*, miRNA expression extends adaxially and is at a higher level than in the wild type. It is not observed in the meristem (arrow). **d**, In 7-day-old *ago1-10*, *PHB* is expressed throughout the leaf primordia. **e**, premir165a and premir165b loci. Orange boxes indicate areas of conserved sequence. Red arrows indicate the position of the miRNA and black arrows the primers used to generate the antisense probe. **f**, Accumulation of transcripts in the P2 leaf of the wild type (1, 2) and *ago1-10* (3, 4). In the wild type, miR165 (red) is expressed abaxially. This leads to accumulation of *PHB* transcript (purple) on the adaxial side of the primordia only (2). In *ago1-10*, miR165 (red) accumulates ectopically on the adaxial side of the primordia (3), resulting in co-localization with *PNH/ZLL* (yellow stipple). Loss of miRNA-directed regulation in *ago1-10* results in ectopic accumulation of *PHB* transcript (4) and an adaxialized phenotype.

abaxial cell fate in *ago1* indicates that miRNA regulation is required for leaf polarity. Other factors may guide the polarizing signal, but this remains to be determined. □

Methods

Growth conditions and plant material

All mutant and transgenic lines were analysed in the Landsberg *erecta* background, unless otherwise noted. *ago1-10* and *ago1-11* arose as point mutants in our gene-trap populations. *ago1-12* was kindly provided by S. Poethig, *dcl1-9* by S. Jacobsen, *rev-6* by L. Comai, and *pnh-2* by K. Barton. Gene-trap and enhancer-trap lines were generated as previously described³⁰.

Seedlings were grown on germination medium containing 1 × MS salts (Gibco), 1% sucrose and 0.5% phytagel (Sigma). Soil-grown plants were grown in Metromix 200 (Scotts) supplemented with 14–14-14 Osmocote (Scotts) at a rate of 2.65 kg m⁻³ and Marathon systemic insecticide (Olympic) at a rate of 0.88 kg m⁻³. Plants were grown at 22 °C in 24 h light (200 microeinsteins m⁻² s⁻¹). *dcl1-9/+*, *ET2689/+*, *ET3964/+* seedlings were selected on germination medium containing kanamycin (50 µg ml⁻¹) and 0.7% agar instead of phytagel.

Histochemical localization of GUS activity

GUS staining was carried out using a substrate solution containing 100 mM sodium phosphate pH 7, 10 mM EDTA, 0.1% Triton X-100, 0.5 µg ml⁻¹ 5-bromo-4-chloro-3-indolyl-β-D-glucuronic acid (X-Gluc), 100 µg ml⁻¹ chloramphenicol, and 2 mM each of potassium ferricyanide and potassium ferrocyanide.

Molecular biology

DNA and RNA extraction and manipulation were carried out using standard methods. Polymerase chain reaction with reverse transcription (RT-PCR) was carried out using total RNA from 21-day-old plants and the Qiagen Onestep kit. Products were sequenced or probed with complementary DNAs to confirm their identity: for *FILAMENTOUS FLOWER* (*FIL*) the oligonucleotide primers were 5'-GTTTGTTCCACCGGACCACTT-3' forward and 5'-TTCTTGGCAGCAGCACTAA-3' reverse; for *KANADI1* (*KAN1*) the primers were 5'-GGAGGAGGAGACGTGGATCA-3' forward and 5'-ACCTGCAATGGCTCTTCAC-3' reverse; for premiR165 the primers were 5'-CCCATCATTCCTCAT CATAA-3' forward and 5'-AAGCCTGGTCCGACGATAC-3' reverse; for *PHB-3* the primers were 5'-GCCAAGGCATCTATGTTGCT-3' forward and 5'-AATGTGAAAACCGGTGAAGC-3' reverse.

In situ hybridizations were performed using a protocol from J. Long (www.its.caltech.edu/~plantlab/protocols/insitu.pdf), using embryos provided by K. Barton and 7-day-old wild-type and *ago1-10* seedlings. premiR165 RT-PCR and PHB-3' products were cloned into pCR2.0 to make RNA probes. *STM* was used as a positive control and a sense transcript from the premiR165b locus that did not include the miRNA sequence was used as a negative control.

Received 6 August 2003; accepted 23 January 2004; doi:10.1038/nature02366.

- Rhoades, M. W. *et al.* Prediction of plant microRNA targets. *Cell* **110**, 513–520 (2002).
- McConnell, J. R. *et al.* Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. *Nature* **411**, 709–713 (2001).
- McConnell, J. R. & Barton, M. K. Leaf polarity and meristem formation in *Arabidopsis*. *Development* **125**, 2935–2942 (1998).
- Sussex, I. M. Morphogenesis in *Solanum tuberosum* L.: Experimental investigation of leaf dorsiventrality and orientation in the juvenile shoot. *Phytomorphology* **5**, 286–300 (1955).
- Bohmert, K. *et al.* AGO1 defines a novel locus of *Arabidopsis* controlling leaf development. *EMBO J.* **17**, 170–180 (1998).
- Lynn, K. *et al.* The PINHEAD/ZWILLE gene acts pleiotropically in *Arabidopsis* development and has overlapping functions with the ARGONAUTE1 gene. *Development* **126**, 469–481 (1999).
- Moussian, B., Schoof, H., Haecker, A., Jurgens, G. & Laux, T. Role of the ZWILLE gene in the regulation of central shoot meristem cell fate during *Arabidopsis* embryogenesis. *EMBO J.* **17**, 1799–1809 (1998).
- Carmell, M. A., Xuan, Z., Zhang, M. Q. & Hannon, G. J. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.* **16**, 2733–2742 (2002).
- Kataoka, Y., Takeichi, M. & Uemura, T. Developmental roles and molecular characterization of a *Drosophila* homologue of *Arabidopsis* Argonaute1, the founder of a novel gene superfamily. *Genes Cells* **6**, 313–325 (2001).
- Harris, A. N. & Macdonald, P. M. *Aubergine* encodes a *Drosophila* polar granule component required for pole cell formation and related to eIF2C. *Development* **128**, 2823–2832 (2001).
- Williams, R. W. & Rubin, G. M. ARGONAUTE1 is required for efficient RNA interference in *Drosophila* embryos. *Proc. Natl Acad. Sci. USA* **99**, 6889–6894 (2002).
- Caudy, A. A., Myers, M., Hannon, G. J. & Hammond, S. M. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes Dev.* **16**, 2491–2496 (2002).
- Mourelatos, Z. *et al.* miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728 (2002).
- Ilavce, C., Kasschau, K. D., Rector, M. A. & Carrington, J. C. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**, 1605–1619 (2002).
- Cubas, P., Lauter, N., Doebley, J. & Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *Plant J.* **18**, 215–222 (1999).
- Palatnik, J. F. *et al.* Control of leaf morphogenesis by microRNAs. *Nature* **425**, 257–263 (2003).
- Eshed, Y., Baum, S. F., Perea, J. V. & Bowman, J. L. Establishment of polarity in lateral organs of plants. *Curr. Biol.* **11**, 1251–1260 (2001).
- Kerstetter, R., Bollman, K., Taylor, R., Bombles, K. & Poethig, R. KANADI regulates organ polarity in *Arabidopsis*. *Nature* **411**, 706–709 (2001).

- Sawa, S. *et al.* *FILAMENTOUS FLOWER*, a meristem and organ identity gene of *Arabidopsis*, encodes a protein with a zinc finger and HMG-related domains. *Genes Dev.* **13**, 1079–1088 (1999).
- Siegrfried, K. R. *et al.* Members of the YABBY gene family specify abaxial cell fate in *Arabidopsis*. *Development* **126**, 4117–4128 (1999).
- The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Emery, J. F. *et al.* Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. *Curr. Biol.* **13**, 1768–1774 (2003).
- Otsuga, D., DeGuzman, B., Prigge, M. J., Drews, G. N. & Clark, S. E. *REVOLUTA* regulates meristem initiation at lateral positions. *Plant J.* **25**, 223–236 (2001).
- Reinhart, B. J. & Bartel, D. P. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**, 1831 (2002).
- Tang, G., Reinhart, B. J., Bartel, D. P. & Zamore, P. D. A biochemical framework for RNA silencing in plants. *Genes Dev.* **17**, 49–63 (2003).
- Chen, Q. *et al.* The *Arabidopsis* *FILAMENTOUS FLOWER* gene is required for flower formation. *Development* **126**, 2715–2726 (1999).
- Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. MicroRNAs in plants. *Genes Dev.* **16**, 1616–1626 (2002).
- Song, J. J. *et al.* The crystal structure of the Argonaute PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nature. Struct. Biol.* **10**, 1026–1032 (2003).
- Foster, T. M. *et al.* A surveillance system regulates selective entry of RNA into the shoot apex. *Plant Cell* **14**, 1497–1508 (2002).
- Sundaresan, V. *et al.* Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9**, 1797–1810 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Barton, B. Reinhart and D. Bartel for sharing unpublished data, M. Timmermans and M. Byrne for discussion and critical reading of the manuscript. We also thank T. Mulligan for plant care, and B. Lehner, S. Gao, S. Morgan, R. Shen, R. Umamaheswari and A. Tang, for laboratory assistance. We thank C. Dean, S. Poethig, J. Levin, R. Howden and J. Moore for alerting us to new alleles of *ago1*. This work was supported by a postdoctoral fellowship from the Cold Spring Harbor Association (C.K.), and grants from the National Science Foundation and the NIH to R.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.M. (martiens@cshl.org).

microRNA-mediated repression of rolled leaf1 specifies maize leaf polarity

Michelle T. Juarez^{1,2}, Jonathan S. Kui^{1,3}, Julie Thomas¹, Bradley A. Heller¹ & Marja C. P. Timmermans¹

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

²Graduate Program in Genetics, State University of New York, Stony Brook, New York 11794, USA

³The Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

In both animals and plants, many developmentally important regulatory genes have complementary microRNAs (miRNAs), which suggests that these miRNAs constitute a class of developmental signalling molecules¹. Leaves of higher plants exhibit a varying degree of asymmetry along the adaxial/abaxial (upper/lower) axis. This asymmetry is specified through the polarized expression of class III homeodomain/leucine zipper (*HD-ZIP III*) genes^{2–4}. In *Arabidopsis*, three such genes, *PHABULOSA* (*PHB*), *PHAVOLUTA* (*PHV*) and *REVOLUTA* (*REV*), are expressed throughout the incipient leaf, but become adaxially localized after primordium emergence. Downregulation of the *HD-ZIP III* genes allows expression of the *KANADI* and *YABBY* genes, which specify abaxial fate^{5–8}. *PHB*, *PHV* and *REV* transcripts contain a complementary site for miRNA165 and miRNA166, which can direct their cleavage *in vitro*^{9–11}. Here we show that miRNA166 constitutes a highly conserved polarizing signal whose

expression pattern spatially defines the expression domain of the maize *hd-zipIII* family member *rolled leaf1* (*rld1*). Moreover, the progressively expanding expression pattern of miRNA166 during leaf development and its accumulation in phloem suggests that miRNA166 may form a movable signal that emanates from a signalling centre below the incipient leaf.

Disruption of the miRNA165/166 complementary site, as in the gain-of-function *phb-d* and *phv-d* alleles, leads to persistence of mutant transcripts in the abaxial domain and formation of adaxialized leaves^{2,12}. Similar dominant mutations in *REV* cause vascular patterning defects, suggesting that specification of adaxial (xylem) and abaxial (phloem) cell fates during vascular development may be governed by a mechanism similar to that underlying the patterning of lateral organs³. The semi-dominant *Rolled leaf1-Original* (*Rld1-O*) mutant causes an upward curling of the leaf blade caused by adaxialization or partial reversal of leaf polarity¹³ (Fig. 1; M.T.J., R. Twigg and M.C.P.T., unpublished results). Early primordium development, including founder-cell recruitment and formation of major lateral veins, occurs normally in *Rld1-O*. However, adaxialized sectors arise on either side of the midvein in older *Rld1-O* primordia. Such sectors differentiate adaxial epidermal characters, including the ligule, on the abaxial side and frequently lack minor vascular bundles and photosynthetic cell types (Fig. 1c).

The *Rld1-O* defects resemble the *phb-d*, *phv-d* and *rev-d* mutant phenotypes in *Arabidopsis*. We therefore characterized several *hd-zipIII* family members from maize. A partial complementary DNA clone with high homology to the carboxy terminus of *REV* was mapped within 0.11 cM of *rld1* (Fig. 2a; see Methods). Characterization of full-length cDNA clones revealed that the maize *rev1* homologue encodes an 840-amino-acid protein with a high degree of sequence identity (70%) to *Arabidopsis* *REV* (Fig. 2b). The miRNA165/166 complementary site^{9,10} is conserved between *Arabidopsis* and maize. *Arabidopsis* *REV*, *PHB* and *PHV* contain an intron at position 9 of the miRNA165/166 complementary site and the position of that intron is also conserved in maize *rev1* (Fig. 2a). Maize and *Arabidopsis* last shared a common ancestor over 70 million years ago, so this conservation suggests an important role in plant development for *REV* and for the potential spatial regulation of *REV* by miRNAs.

Two dominant *Rld1* alleles, *Rld1-5* and *Rld1-6*, arose spontaneously in *Mutator*-active families. A single nucleotide change in the 5' end of the miRNA165/166 complementary site of *rev1*, which causes a G-to-D amino-acid substitution, distinguishes *Rld1-5* and *Rld1-6* from their respective progenitor alleles (Fig. 2b, c). *Rld1-O* and *Rld1-N1441*, which were identified previously in chemically mutagenized populations¹³, have the identical

nucleotide change in the miRNA165/166 complementary site. Seven of the nine dominant *phb* and *phv* mutations also map to a single nucleotide, albeit near the 3' end of the miRNA165/166 complementary site². Together, these observations suggest that *rev1* is *rld1* and that the G-to-D amino-acid substitution in the START lipid-sterol binding domain or misexpression of *rev1* leads to the *Rld1* mutant defects.

rld1 is normally expressed at the tip of the shoot apical meristem (SAM) and in a stripe from the centre of the SAM to the site of leaf initiation (Fig. 3a). In the P1 primordium, *rld1* is expressed along the adaxial domain as well as in the midvein region (Fig. 3a, b). *rld1* expression persists during primordium development in the vasculature and on the adaxial side near the margins. This expression pattern resembles that of *HD-ZIPIII* genes in *Arabidopsis*²⁻⁴ and suggests a role for *rld1* in meristem function and adaxial cell fate determination. The meristematic expression pattern of *rld1* is unaffected in *Rld1-O*. However, *rld1* is misexpressed on the abaxial side of *Rld1-O* leaf primordia and persists in a broader domain during *Rld1-O* primordium development (Fig. 3c). Although strong *rld1* expression does become restricted to the margins and vasculature, uniform weak expression is observed throughout the central domain of P2 *Rld1-O* primordia. Moreover, the domain of expression near the margins is broader and includes the abaxial side. In subsequently older *Rld1-O* primordia, *rld1* expression is gradually reduced in the central region.

To determine whether adaxialization of *Rld1-O* leaf primordia is

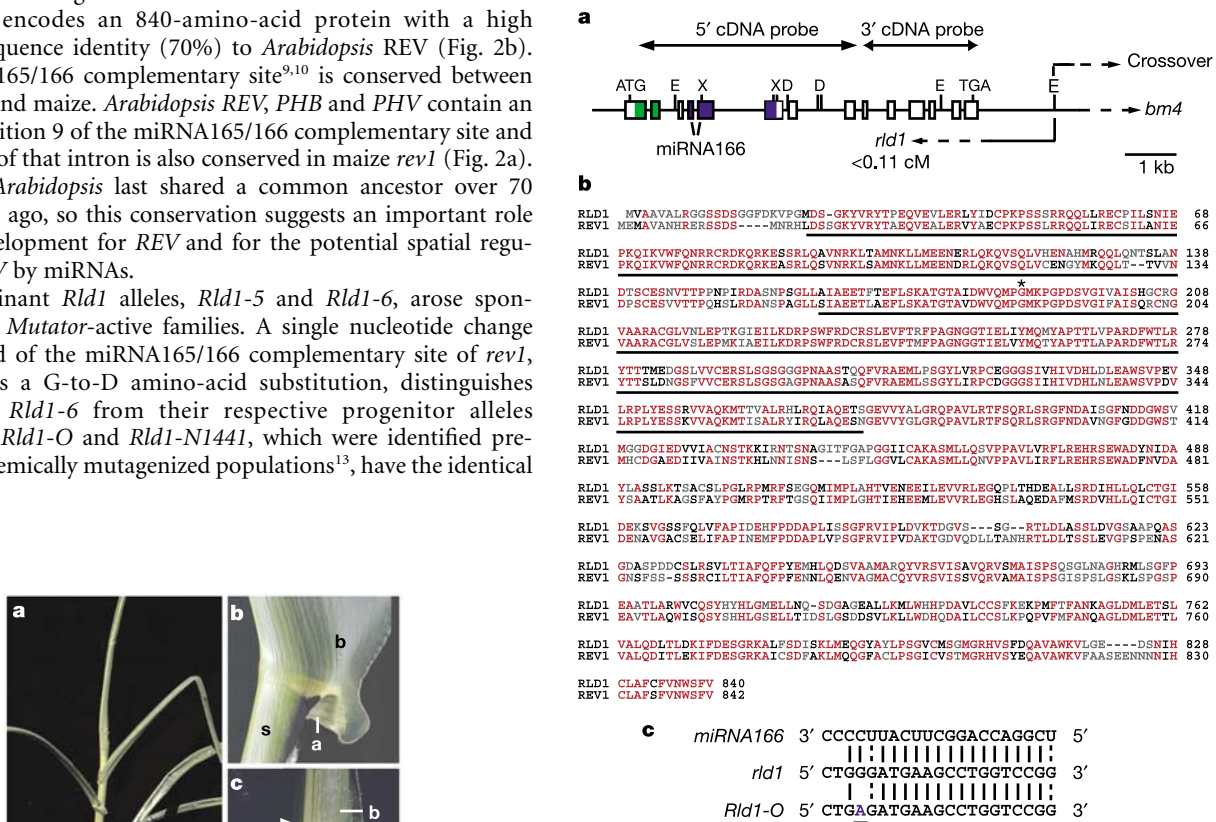


Figure 1 Effects of *Rld1-O* on leaf polarity. **a**, Mature *Rld1-O* plant. **b**, **c**, Abaxial surfaces of wild-type (**b**) and *Rld1-O* (**c**) adult leaves. Sheath (s) and blade (b) are separated by auricle (a). The black arrow marks the ectopic ligule; the white arrow marks an adaxialized blade sector.

Figure 2 Dominant *Rld1* mutations affect the miRNA166 complementary site. **a**, Organization of the *rld1* gene. Boxes, exons; green, HD-ZIP domain; blue, START domain. The miRNA166 complementary site, relevant probes and restriction enzymes are indicated: D, *Dra*I; E, *Eco*RV; X, *Xba*I. **b**, Alignment of the maize RLD1 and *Arabidopsis* REV proteins. Red, identity; black, similarity. The HD-ZIP and START domains are underlined. Asterisk marks the mutation in dominant *Rld1* alleles. **c**, Alignment of miRNA166 with the target sites in *rld1* and *Rld1-O*. Solid lines, Watson-Crick base pairs; dotted lines, RNA base pairs. The mutation in *Rld1-O* is highlighted in blue.

correlated with misexpression of *rld1* specifically, we compared the expression patterns of a maize *PHB* homologue in wild-type and *Rld1-O* mutant apices. *phb* expression in the wild type occurs in a pattern comparable to that of *rld1*, although *phb* expression is relatively more abundant in developing vasculature (Fig. 3d). The overlapping expression patterns of *rld1* and *phb* suggest that these genes may have redundant functions. The *phb* expression pattern is unaffected in *Rld1-O* (Fig. 3e). These results together with sequence analysis of the dominant *Rld1* alleles indicate that miRNAs may mediate the post-transcriptional repression of *rld1* on the abaxial side of leaf primordia. However, the lack of *rld1* expression in the central region of older *Rld1-O* primordia indicates that the *rld1* expression domain is also controlled in part at the transcriptional level.

miRNAs are processed from larger imperfect hairpin precursors (pre-miRNA) by the double-stranded RNase *DICER-LIKE1* (refs. 9,14,15). *Arabidopsis* contains two potential pre-miRNA165 (*MIR165*) and seven *MIR166* loci⁹. Pairwise comparisons between

MIR166 family members revealed extensive sequence conservation outside the predicted hairpin structures (Fig. 4a; data not shown). *MIR166a* and *MIR166b* share nine conserved sequence motifs over a region of ~1.5 kilobases (kb), which range in size from 28 to 122 base pairs (bp). The rice genomic sequence includes at least six loci with the potential to generate miRNA166 homologues although *MIR165* loci have not been identified⁹. Over a 392-bp region, *OsMIR166b* and *OsMIR166d* share three conserved domains, including the miRNA and fold-back sequences, that range in size from 28 to 47 bp. Interestingly, the sequence and arrangement of these three motifs is conserved in *Arabidopsis MIR166a* and

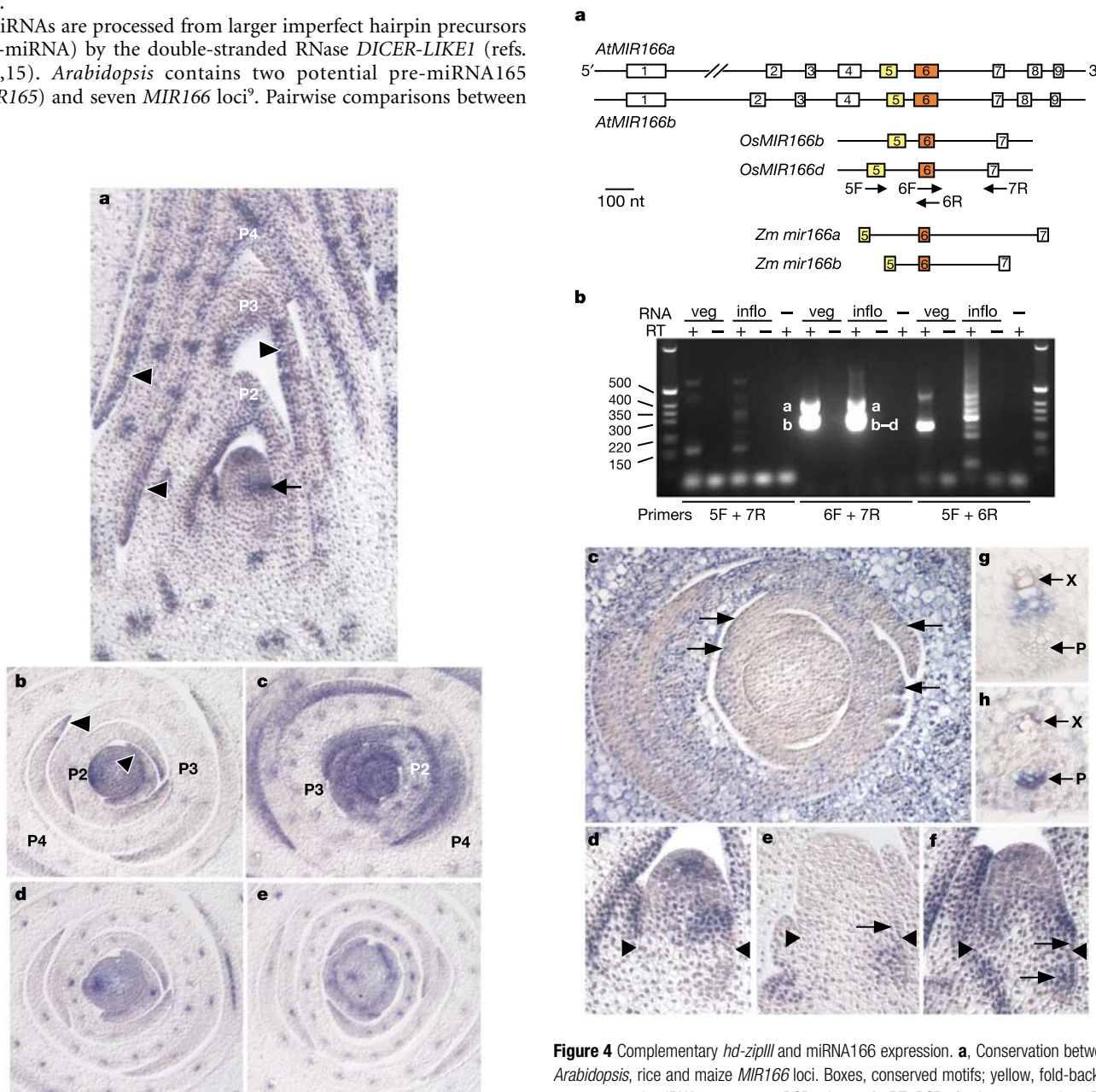


Figure 4 Complementary *hd-zipIII* and miRNA166 expression. **a**, Conservation between *Arabidopsis*, rice and maize *MIR166* loci. Boxes, conserved motifs; yellow, fold-back sequence; red, miRNA166; arrows, PCR primers. **b**, RT-PCR of *mir166* transcripts. RNA sources, primer pairs and inclusion of reverse transcriptase (RT) are indicated. Labels a, b and b-d indicate *mir166a-d*-derived products, respectively. **c**, miRNA166 expression in wild-type leaf primordia. Arrows, abaxial expression. **d, e**, *rld1* (d) or miRNA166 (e) expression in wild-type apices. Arrowheads, top P1 disk of insertion; arrow, abaxial miRNA166 expression in the incipient leaf. **f**, *rld1* misexpression in *Rld1-O* near the incipient leaf (arrows). **g, h**, *phb* (g) or miRNA166 (h) expression in wild-type vascular bundles. X, xylem; P, phloem.

MIR166b (Fig. 4a), suggesting that they constitute part of the pre-miRNA transcript or contain elements important for miRNA166 regulation that may also be conserved in maize. Long precursor transcripts have recently been identified for *MIR-JAW* and *MIR172a*^{16,17}.

Degenerate primers derived from the conserved sequence motifs (Fig. 4a) allowed amplification of reverse-transcriptase-dependent products from maize vegetative apex and immature tassel RNA, although not with equal efficiency (Fig. 4b). Sequence analysis suggests at least four *mir166* genes (*mir166a–mir166d*) that contain the identified conserved motifs are present in maize. No primary sequence similarity was observed outside these motifs among the maize genes and between the maize, rice and *Arabidopsis* genes. Consistent with the amplification of distinct polymerase chain reaction (PCR) products from vegetative and inflorescence tissues, all four *mir166* genes are expressed in immature tassels, whereas only *mir166a* and *mir166b* were found to be expressed in vegetative apices (Fig. 4b). The accumulation of *Arabidopsis* miRNAs is developmentally regulated^{9,14–17}, but these results suggest that individual pre-miRNA genes can display distinct tissue specificities. Sequence analysis of partial genomic clones for *mir166a* and *mir166b* suggests that these genes produce both unspliced and spliced transcripts, similar to *MIR172a* from *Arabidopsis*¹⁶.

miRNAs have been suggested to control a variety of developmental processes in both animals and plants¹. Characterization of the dominant *Rld1*, *phb*, *phv* and *rev* alleles^{2,3} implicate, miRNA165/166 in abaxial cell fate specification by restricting *hd-zipIII* expression to the adaxial side. To examine directly whether miRNAs can establish patterns of tissue differentiation during development, we used *in situ* hybridization to compare the *rld1* and miRNA166 expression patterns at the cellular level. Consistent with the relatively low abundance of precursor transcripts^{9,14–17}, no hybridization was detected using probes excluding the miRNA (data not shown). *In situ* hybridization using a fragment of *mir166a* including the miRNA revealed expression in developing leaf primordia but not in the SAM (Fig. 4c, e). *rld1* is expressed in a stripe of cells at the site of leaf initiation (Fig. 4d). miRNA166 accumulates below the incipient leaf opposite the P1 primordium, although weak miRNA166 expression can occur on the abaxial side of the P0 leaf (Fig. 4e). In the P1 primordium, high levels of miRNA166 accumulate on the abaxial side in a slightly more restricted pattern than defined by the lack of *rld1* (Fig. 4c), and low levels of miRNA166 may accumulate in a slightly broader domain. miRNA166 is also expressed below the P1 leaf (Fig. 4e). In older leaf primordia, miRNA166 accumulates in a progressively broader domain extending laterally and adaxially. Only near the margins does miRNA166 expression remain limited to the abaxial side (Fig. 4c).

rld1 and miRNA166 thus exhibit complementary expression patterns in developing leaf primordia, consistent with a role for miRNA166 in the spatial regulation of *rld1* transcripts. However, the lack of miRNA166 expression in the SAM suggests that the meristematic *rld1* expression domain is controlled at the transcriptional level. Consistent with this possibility, *rld1* is ectopically expressed in *Rld1-O* and includes the miRNA166 expression domain below the incipient leaf (Fig. 4f), but *rld1* expression in *Rld1-O* SAMs is unaffected. The miRNA166 expression pattern is unaltered in *Rld1-O*, suggesting that the *mir166* genes act upstream of *rld1* (data not shown). *hd-zipIII* genes also function in adaxial/abaxial patterning of vascular bundles^{3,18,19}. Vascular *phb* expression is indeed limited to the adaxial pro-xylem cells (Fig. 4g). In contrast, miRNA166 accumulates in the abaxial phloem tissue (Fig. 4h), supporting conservation between adaxial/abaxial patterning during vascular and lateral organ development.

Mutations that disrupt complementarity to the 5' end of miRNAs are known to affect miRNA-mediated cleavage^{2,3,11,17,20}. Characterization of *Rld1* indicates that nucleotide substitutions disrupting complementation at the miRNA 3' end similarly interfere with

miRNA function. *Rld1* transcripts accumulate on the abaxial side in initiating and young *Rld1-O* leaf primordia but this is not sufficient to prevent initial adaxial/abaxial patterning. Lateral founder-cell recruitment and early vascular development are unaffected in *Rld1-O* primordia and these processes depend on specification of adaxial/abaxial polarity²¹. Remaining differential *hd-zipIII* expression during early primordium development may initially facilitate normal adaxial/abaxial patterning in *Rld1-O*. Alternatively, RLD1 or other components of the adaxial specification pathway may need to be activated to adaxialize *Rld1-O* primordia. Specification of adaxial cell fate is thought to require a meristem-born signal because separation of incipient primordia from the SAM by incision causes formation of radially symmetric abaxialized leaves^{22,23}.

Despite the differences in monocot and dicot leaf development, the spatial regulation of *hd-zipIII* genes is essential for adaxial/abaxial patterning in both classes of angiosperms, indicating that miRNA165/166 constitutes an important, highly conserved polarizing signal. miRNA166 initially accumulates immediately below the incipient leaf but subsequently in a progressively broader domain including the adaxial side. This dynamic expression pattern and the possible gradient of miRNA166 expression in leaf primordia is reminiscent of a movable signal, suggesting that expression of the miRNA166 precursor could be under such control or alternatively that miRNA166 can move between cells. Although our data are consistent with both possibilities, the observation that miRNA166 accumulates in phloem leads us to propose that miRNA166 may move from the site of *mir166* expression. Consistent with this possibility, expression of viral movement proteins that disrupt RNA trafficking causes formation of radial adaxialized leaves²⁴. The site of *mir166* expression is unknown but may coincide with the initial strong miRNA166 accumulation below the incipient primordium. If so, the *mir166* expression domain may be established independently of abaxial determinants that function in the incipient primordium^{5–8}, and may constitute a signalling centre for polarity in the leaf. □

Methods

Genetic analysis

rev1 was mapped near *rld1* on chromosome arm 9L using recombinant inbred populations². Recombinants between *Rld1-O* and *brown midrib4* (*bm4*) were generated to determine a precise map position for *rev1*. *Rld1-O/+bm4* double heterozygotes were test-crossed onto *+bm4/bm4* plants. One *+bm4/Rld1-O bm4* and four *+bm4/+bm4* recombinants were identified among 883 plants, suggesting a map distance of 0.57 cM between *rld1* and *bm4*. DNA from five recombinants and from pools of non-recombinant siblings was digested with different restriction enzymes and hybridized with the 3' *rev1* cDNA probe. Two recombinants between *rev1* and *rld1* were identified upon digestion with *XbaI* and *DraI*, suggesting that the 3' end of *rev1* maps between *rld1* and *bm4*. However, no recombinants were identified upon digestion with *EcoRV*. This suggests that both crossovers occurred within the *XbaI* and *DraI* genomic fragments that hybridized to the probe, but downstream of both *EcoRV* sites (Fig. 2a) (for explanation see ref. 26). As no recombination was observed between *rev1* and *rld1* upon *EcoRV* digestion, *rld1* maps within 0.11 cM upstream of the 3' *EcoRV* site. Consistent with this data, no recombinants were subsequently observed between the 5' end of *rev1* and *rld1* with all restriction enzymes tested.

Molecular biology

RNA was prepared using standard protocols from vegetative apices including the SAM and five leaf primordia and from immature tassels (1–2 cm) that include developing flowers. Full-length *rld1* cDNA clones were obtained by reverse transcription (RT)–PCR using standard protocols and primers based on sequence conservation between the *Arabidopsis* and rice *REV* genes (CNAAGCAGATCAANGTCTGG) or based on maize genome survey sequencing reads (GAGAGCTAAGAGCAACAAGG). *rld1* genomic clones were assembled using maize genome survey sequencing data and standard PCR and sequencing protocols. A partial *phb* cDNA clone was isolated by RT–PCR using gene-specific primers based on available expressed sequence tag (EST) sequence data (GTGCCGAATGCTGCTTCAG and TGGATCCGCTCTTGATCC). Partial *MIR166* cDNA and genomic clones were amplified from maize and *Arabidopsis* using the following primers. 5F, TGAGNGGAATGTTGTCTGGY; 6F, GNTCTCGGACCAGGCTTCA; 6R, GAATGAAGCCTGGTCCGAG; 7R, NNYCATSATTACCAATCTG. 6F + 7R RT–PCR products and 5F + 7R genomic fragments were cloned and sequenced.

Sequence analysis

Protein or DNA sequences were aligned using ClustalW (MacVector 6.5.3). Sequences

extending 2 kb on either side of miRNA166 were identified from GenBank by BLAST analyses. Pairwise comparisons between *MIR166* genes were made using Pustell DNA matrix analysis (MacVector 6.5.3). GenBank accession numbers are: *rdl1*, AY501430; *mir166a*, AY501431; *mir166b*, AY501432; *mir166c*, AY501433; *mir166d*, AY501434.

In situ hybridization

Tissue sections prepared from shoot apices of two-week-old mutant and wild-type seedlings were pretreated and hybridized as described²⁷. Digoxigenin-labelled probes were prepared by *in vitro* transcription (Stratagene) according to the manufacturer's protocol. An *rdl1*-specific probe encompassing nucleotides 619–1,674 of the coding sequence, and a *phb*-specific probe encompassing amino acids 462–606 of the *Arabidopsis* PHB protein were used at a concentration of 0.5 ng μl^{-1} kb⁻¹ probe complexity. miRNA166 expression was determined using a *mir166a*-derived probe amplified with primers 6F and 7R. A precursor-specific probe was amplified using primer 7R and a gene-specific primer (CCTCCATCAGATGAGCTCC) downstream of miRNA166.

Received 4 December 2003; accepted 23 January 2004; doi:10.1038/nature02363.

- Carrington, J. C. & Ambros, V. Role of microRNAs in plant and animal development. *Science* **301**, 336–338 (2003).
- McConnell, J. R. *et al.* Role of *PHABULOSA* and *PHAVOLUTA* in determining radial patterning in shoots. *Nature* **411**, 709–713 (2001).
- Emery, J. F. *et al.* Radial patterning of *Arabidopsis* shoots by class III *HD-ZIP* and *KANADI* genes. *Curr. Biol.* **13**, 1768–1774 (2003).
- Otsuga, D., DeGuzman, B., Prigge, M. J., Drews, G. N. & Clark, S. E. *REVOLUTA* regulates meristem initiation at lateral positions. *Plant J.* **25**, 223–236 (2001).
- Eshed, Y., Baum, S. F., Perea, J. V. & Bowman, J. L. Establishment of polarity in lateral organs of plants. *Curr. Biol.* **11**, 1251–1260 (2001).
- Kerstetter, R. A., Bollman, K., Taylor, R. A., Bombli, K. & Poethig, R. S. *KANADI* regulates organ polarity in *Arabidopsis*. *Nature* **411**, 706–709 (2001).
- Sawa, S. *et al.* *FILAMENTOUS FLOWER*, a meristem and organ identity gene of *Arabidopsis*, encodes a protein with a zinc finger and HMG-related domains. *Genes Dev.* **13**, 1079–1088 (1999).
- Siegrfried, K. R. *et al.* Members of the *YABBY* gene family specify abaxial cell fate in *Arabidopsis*. *Development* **126**, 4117–4128 (1999).
- Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. MicroRNAs in plants. *Genes Dev.* **16**, 1616–1626 (2002).
- Rhoades, M. W. *et al.* Prediction of plant microRNA targets. *Cell* **110**, 513–520 (2002).
- Tang, G., Reinhart, B. J., Bartel, D. P. & Zamore, P. D. A biochemical framework for RNA silencing in plants. *Genes Dev.* **17**, 49–63 (2003).
- McConnell, J. R. & Barton, M. K. Leaf polarity and meristem formation in *Arabidopsis*. *Development* **125**, 2935–2942 (1998).
- Nelson, J. M., Lane, B. & Freeling, M. Expression of a mutant maize gene in the ventral leaf epidermis is sufficient to signal a switch of the leaf's dorsoventral axis. *Development* **129**, 4581–4589 (2002).
- Llave, C., Kasschau, K. D., Rector, M. A. & Carrington, J. C. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**, 1605–1619 (2002).
- Park, W., Li, J., Song, R., Messing, J. & Chen, X. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**, 1484–1495 (2002).
- Aukerman, M. J. & Sakai, H. Regulation of flowering time and floral organ identity by a microRNA and its *APETALA2*-like target genes. *Plant Cell* **15**, 2730–2741 (2003).
- Palatnik, J. F. *et al.* Control of leaf morphogenesis by microRNAs. *Nature* **425**, 257–263 (2003).
- Zhong, R. & Ye, Z. H. *IFL1*, a gene regulating interfascicular fiber differentiation in *Arabidopsis*, encodes a homeodomain-leucine zipper protein. *Plant Cell* **11**, 2139–2152 (1999).
- Ratcliffe, O. J., Riechmann, J. L. & Zhang, J. Z. *INTERFASCICULAR FIBERLESS1* is the same gene as *REVOLUTA*. *Plant Cell* **12**, 315–317 (2000).
- Llave, C., Xie, Z., Kasschau, K. D. & Carrington, J. C. Cleavage of *Scarecrow*-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**, 2053–2056 (2002).
- Timmermans, M. C. P., Schultes, N. P., Jankovsky, J. P. & Nelson, T. *Leafbladeless1* is required for dorsoventrality of lateral organs in maize. *Development* **125**, 2813–2823 (1998).
- Sussex, I. M. Morphogenesis in *Solanum tuberosum* L.: Experimental investigation of leaf dorsoventrality and orientation in the juvenile shoot. *Phytomorphology* **5**, 286–300 (1955).
- Snow, M. & Snow, R. The dorsoventrality of leaf primordia. *New Phytol.* **58**, 188–207 (1959).
- Foster, T. M. *et al.* A surveillance system regulates selective entry of RNA into the shoot apex. *Plant Cell* **14**, 1479–1508 (2002).
- Burr, B., Burr, F. A., Thompson, K. H., Albertson, M. C. & Stuber, C. W. Gene mapping with recombinant inbreds in maize. *Genetics* **118**, 519–526 (1988).
- Timmermans, M. C. P., Das, O. P. & Messing, J. Characterization of a meiotic crossover in maize identified by a restriction fragment length polymorphism-based method. *Genetics* **143**, 1771–1783 (1996).
- Jackson, D. In *Molecular Plant Pathology: A Practical Approach* (eds Bowles, D. J., Gurr, S. J. & McPherson, M.) 163–174 (Oxford Univ. Press, Oxford, 1991).

Acknowledgements We thank T. Phelps-Durr, C. Kidner, T. Volpe and G. Hannon for discussions and comments on the manuscript. We also thank T. Mulligan for plant care. This work was supported by grants from the USDA and the NSF (to M.C.P.T.). M.T.J. is funded in part by a W. Burghardt Turner Fellowship, and J.S.K. is an Alfred Hershey Fellow of the Watson School of Biological Sciences and is supported by a grant from the National Institute of General Medical Sciences, NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.C.P.T. (timmerma@cshl.org).

A non-B-DNA structure at the *Bcl-2* major breakpoint region is cleaved by the RAG complex

Sathees C. Raghavan^{1,2}, Patrick C. Swanson⁶, Xiantuo Wu¹, Chih-Lin Hsieh^{1,3,4} & Michael R. Lieber^{1,2,3,5}

¹Norris Comprehensive Cancer Center, Room 5428, ²Departments of Pathology, ³Biochemistry & Molecular Biology, ⁴Urology, ⁵Biological Science and Molecular Microbiology & Immunology, University of Southern California Keck School of Medicine, 1441 Eastlake Ave., MC9176, Los Angeles, California 90033, USA ⁶Department of Medical Microbiology and Immunology, Creighton University School of Medicine, 2500 California Plaza, Omaha, Nebraska 68178, USA

The causes of spontaneous chromosomal translocations in somatic cells of biological organisms are largely unknown, although double-strand DNA breaks are required in all proposed mechanisms^{1–5}. The most common chromosomal abnormality in human cancer is the reciprocal translocation between chromosomes 14 and 18 (t(14;18)), which occurs in follicular lymphomas. The break at the immunoglobulin heavy-chain locus on chromosome 14 is an interruption of the normal V(D)J recombination process. But the breakage on chromosome 18, at the *Bcl-2* gene, occurs within a confined 150-base-pair region (the major breakpoint region or Mbr) for reasons that have remained enigmatic. We have reproduced key features of the translocation process on an episome that propagates in human cells. The RAG complex—which is the normal enzyme for DNA cleavage at V, D or J segments—nicks the *Bcl-2* Mbr *in vitro* and *in vivo* in a manner that reflects the pattern of the chromosomal translocations; however, the Mbr is not a V(D)J recombination signal. Rather the *Bcl-2* Mbr assumes a non-B-form DNA structure within the chromosomes of human cells at 20–30% of alleles. Purified DNA assuming this structure contains stable regions of single-strandedness, which correspond well to the translocation regions in patients. Hence, a stable non-B-DNA structure in the human genome appears to be the basis for the fragility of the *Bcl-2* Mbr, and the RAG complex is able to cleave this structure.

The RAG complex can misrecognize and cleave pseudo-signals at lymphoid translocation breakpoints at LMO2, Ttg-1 and SIL⁶. We tested the RAG-misrecognition hypothesis at the *Bcl-2* Mbr, although there is no apparent V(D)J heptamer/nonamer sequence within it (Supplementary Fig. 1). To test whether the *Bcl-2* sequence serves as a recombination signal, a 300-base-pair (bp) *Bcl-2* sequence (containing the 150-bp Mbr region) was positioned 250-bp upstream or downstream of a single 12- or 23-signal sequence on an episome that can replicate in human cells. We transfected the substrates into the human pre-B-cell line Reh, and recovered them 48 h later (Supplementary Fig. 2).

The boundaries of the recovered recombinant molecules were characterized. None of the breakpoint boundaries was at the 12- or 23-signals; rather, the recombinants were far to one side or the other of each signal and clearly not related to V(D)J recombination. However, a disproportionate number of recombination boundaries were in the *Bcl-2* Mbr (Table 1). The recombination frequency (0.056%) was 160-fold lower in our constructs than in a substrate with a pair of optimal 12/23-signals (recombination frequency, 9%) in Reh cells. The pattern of the recombination breakpoints within the *Bcl-2* Mbr was striking. About 77% (27 out of 35) of the breakpoints were within the Mbr, and another five of the 35 events were immediately outside (within five to seven nucleotides) of the *Bcl-2* Mbr (Table 1). Interestingly, most of these 27 events were within peaks I, II or III of the Mbr (data not shown). The difference between the 77% in the Mbr versus the expected random level of

Table 1 Breakpoint frequencies in plasmids from cells expressing RAG proteins

	DA	DAC _{Total}			(DAC _{Mbr} /DA) × 100	% Events inside Mbr	% Events outside Mbr
		DAC ₂₃	DAC _{Mbr}	DAC _{Not Mbr}			
Reh	48,250	0	27	8	5.6×10^{-2}	77.0% (27/35)*	23.0% (8/35)
293 + RAGs	616,100	0	26	8	4.2×10^{-3}	76.5% (26/34)*	23.5% (8/34)
293	1,148,300	0	14	19	1.2×10^{-3}	42.4% (14/33)*	57.6% (19/33)
293 + Mut RAGs	274,300	0	3	5	1.1×10^{-3}	37.5% (3/8)	62.5% (5/8)
Random distribution	—	—	—	—	—	48.0%	52.0%

The episome (pXW5) was transfected into mammalian cells (either Reh or 293T cells with or without RAG constructs). The DNA was harvested by the rapid alkaline lysis method after 48 h (Reh) or 40 h (293T), and the recombinants were selected on ampicillin–chloramphenicol plates after transformation into *Escherichia coli*. Plasmid DNA was further screened by restriction digestion (*Sall/BglII*) to select clones containing breakpoints at the *Bcl*-2 region (inside Mbr, 150 bp or outside Mbr, 160 bp) and sequenced. The number of replicated molecules is indicated as DA, which stands for *DpnI*-resistant, ampicillin-resistant. DAC_{Total} indicates the total number of ampicillin–chloramphenicol double-resistant transformants (recombinants) obtained. Of this group, DAC₂₃ are recombinants that use the 23-signal sequence for recombination (zero here). DAC_{Mbr} are recombinants with a breakpoint at *Bcl*-2 Mbr. DAC_{Not Mbr} is the recombinants with breakpoints at a 160-bp region that is outside the Mbr. The recombination efficiency of Mbr is calculated by dividing DAC_{Mbr} by DA. The frequency of recombinants obtained is indicated as a percentage (% Events). Actual numbers of recombinant clones obtained of a given type (divided by the total recombinants) are indicated in parentheses. Reh indicates recombinants obtained from Reh cells (which have native RAGs); 293 + RAGs indicates that the cells were also transfected with core RAG expression vectors; 293 indicates that the cells were not transfected with RAG vectors; and 293 + Mut RAGs indicates that cells were also transfected with a mutant RAG expression vector. The statistical analysis used the Fisher's exact test.

**P* = 0.006 for a comparison of the values in rows one and three and in rows two and three of '% Events inside Mbr' column.

48% is highly significant (*P* < 0.006). T-nucleotides (templated nucleotides) were also present at the junctions of these recombinants, a signature feature of t(14;18) in patients (data not shown)⁷.

Given that Reh cells have high endogenous levels of RAG proteins, we wondered whether the *Bcl*-2 Mbr breaks are caused by RAGs. To test the *in vivo* contribution of RAG proteins to Mbr recombination, we used the human cell line 293T, which does not

carry out any detectable V(D)J recombination. We transfected this cell line with a plasmid (pXW5) bearing the *Bcl*-2 Mbr (paired with a single 23-signal) with or without RAG1 and RAG2 expression vectors or with mutant RAG1 and RAG2 expression vectors^{8,9}. We observed that the recombination frequencies are consistently about three- to fourfold higher in the cells transfected with RAGs than in those without RAGs or with mutant RAGs (Table 1). The recombi-

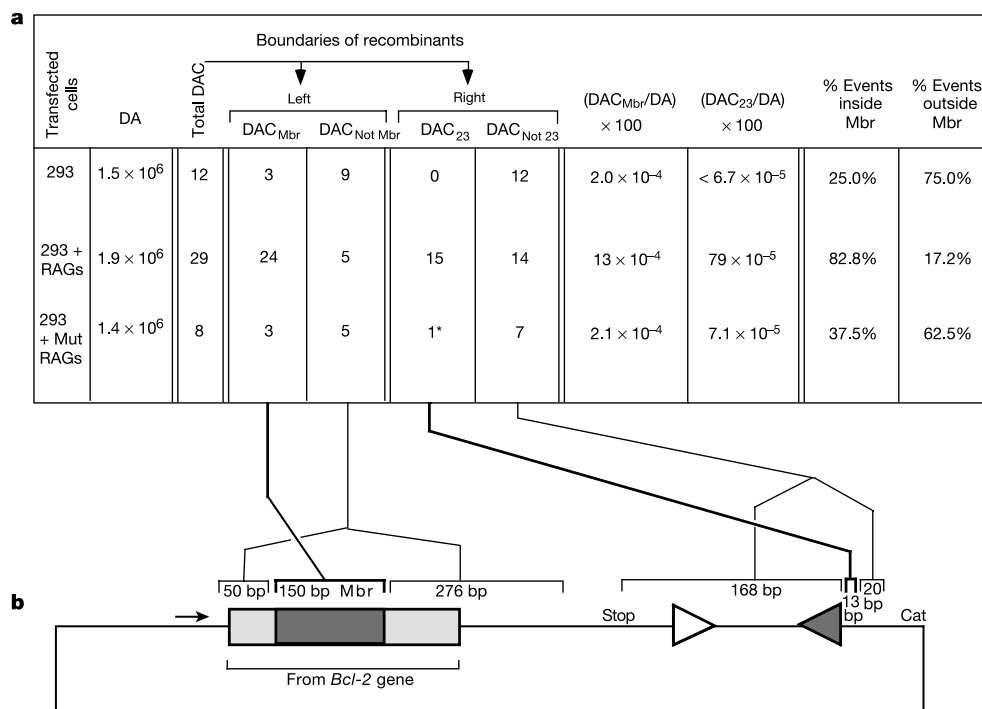


Figure 1 *In vivo* breakpoints on an extrachromosomal substrate cluster within the *Bcl*-2 Mbr and are dependent on the RAG complex. The episome (pSCR45) was transfected into mammalian cells (293T cells), recovered 40 h later and analysed in bacteria as described in the Methods. **a**, Comparison of breakpoint frequencies in recombinant plasmid molecules. DA, the number of substrate molecules that replicated in the 293T cells. Total DAC, bacterial transformants that are ampicillin–chloramphenicol (double) resistant (also referred to as recombinants). The total DAC for each row is subdivided according to the left and right boundaries of recombination. For the left boundary, DAC_{Mbr} is the subset of double-resistant recombinants that have a breakpoint within the *Bcl*-2 Mbr. DAC_{Not Mbr} is the subset of recombinants with breakpoints anywhere within the 326-bp region upstream or downstream of the Mbr. For the right boundary, DAC₂₃ are recombinants that use the 23-signal sequence for recombination. DAC_{Not 23} are recombinants that do not use the 23-signal for recombination. (DAC_{Mbr}/DA) × 100 is the recombination frequency

for all events that use the Mbr. (DAC₂₃/DA) × 100 is the recombination frequency for all events that use the 23-signal. The frequency of recombinants obtained is indicated as a percentage (% Events). The row labelled 293 indicates recombinants obtained from 293T cells in the absence of RAGs; 293 + RAGs indicates that cells were also transfected with full-length RAG expression vectors; and 293 + Mut RAGs indicates that the cells were transfected with mutant RAG1 and full-length RAG-2 vectors. The asterisk indicates that this single event may represent a random break that just happens to be within the 13-bp region adjacent to the heptamer of the 23-signal¹⁰. **b**, Regions of recombination relative to the sequence of pSCR45. The Mbr region is shown in dark grey. The light grey regions are outside of the Mbr but are the flanking regions that naturally are adjacent to the Mbr in the human chromosome. The 12-signal is indicated by the open triangle, and the 23-signal by the filled triangle. The transcriptional promoter (short arrow upstream of the Mbr), the transcription terminator (Stop) and the chloramphenicol gene (Cat) are indicated.

nation frequency of pXW5 (0.004%) in the presence of RAGs is about 25-fold lower than that of a substrate with a pair of optimal 12/23-signals (recombination frequency of 0.1% in fibroblasts transfected with RAG vectors⁹). More importantly, the distribution of the breakpoints was markedly influenced by the presence of the RAG proteins. With RAGs, the breaks were predominantly within the *Bcl-2* Mbr, with particular predilection for peaks I, II and III (76.5%, 26 out of 34 events) (Table 1).

We observed clustering of breaks at the *Bcl-2* Mbr side of the deletion events, but the other boundary of the deletions (recombinants) above were not at the 23-signal. In clinical cases of the *Bcl-2* translocation, breaks at the immunoglobulin heavy chain occur during attempted rearrangement between D and J elements that use 12- and 23-signals (Supplementary Fig. 1). Therefore, we designed a substrate to check the recombination efficiency of the *Bcl-2* Mbr when both the 12- and 23-signals are present rather than one alone (Fig. 1). This plasmid, pSCR45, was then transfected into 293T cells with and without human full-length RAG expression vectors, or with a mutant RAG1 vector¹⁰.

In common with the observations described above for one-signal substrates (Table 1), double-signal substrates have a breakpoint distribution that is strikingly dependent on the Mbr (24 out of 29 events) (Fig. 1a). The preference for breaks in the Mbr occurs despite twice as much DNA surrounding the area where breaks could occur (150 bp in the Mbr compared with 326 bp both upstream and downstream). In more than half of the events (15 out of 29), the second break is at the 23-signal, such that the coding end formerly attached to that signal is now joined to the *Bcl-2* Mbr break. Such a precise use of the 23-signal only in the simultaneous presence of a 12-signal clearly indicates the involvement of the RAG complex. The presence of both a 12- and a 23-signal permits a recapitulation of the D to J joining process, whereas a single signal does not (Table 1; see also Supplementary Fig. 2). This indicates that the *Bcl-2* Mbr recombination with the coding end of the 23-signal is

likely to be an interruption of the normal 12/23-recombination. In the absence of RAGs or with mutant RAGs, there is little or no cleavage at the 23-signal (Fig. 1a).

Given that the *Bcl-2* Mbr recombined with a coding end adjacent to a 23-signal, the involvement of the RAG proteins is clear. But if the RAG complex recombines the Mbr *in vivo*, is it possible that it might cleave the Mbr in a purified system consisting only of the RAG complex and the Mbr DNA? To test this, we incubated the *Bcl-2*-Mbr-bearing plasmids with the purified RAG complex under standard physiological divalent cation (Mg^{2+}) buffer conditions for RAG cleavage reactions. Although the core and full-length RAG complexes do not nick the top strand of the Mbr, they consistently nick the bottom strand to an extent that is similar to its nicking of a standard 23-signal (Supplementary Fig. 3). Active-site mutant RAGs and controls (which do not have RAGs) showed no nicking (see Supplementary text).

These *in vitro* nicking results by RAGs are consistent with the *in vivo* substrate recombination. However, what feature of the Mbr does the RAG complex recognize if the Mbr does not function as a recombination signal?

In the absence of any heptamer/nonamer function by the *Bcl-2* Mbr, we wondered whether there is a structural basis for the fragility at this 150-bp region. To test for single-strandedness of the *Bcl-2* Mbr, we used chemical-probing methods on genomic DNA^{11,12}. To study the single-strandedness of chromosomal DNA at the single-molecule level, bisulphite-treated DNA is polymerase chain reaction (PCR)-amplified, cloned and sequenced. Bisulphite converts unpaired cytosines to uracil, and these become thymine after PCR amplification, thus allowing the detection of single-stranded regions¹¹.

The chromosomal DNA extracted from Reh cells was used for bisulphite treatment. A PCR fragment of 528 bp containing the *Bcl-2* Mbr amplified from treated DNA was cloned, sequenced and analysed. Most cytosine conversions occur in peaks I and III of the Mbr region (Fig. 2). If we display the results for individual strands,

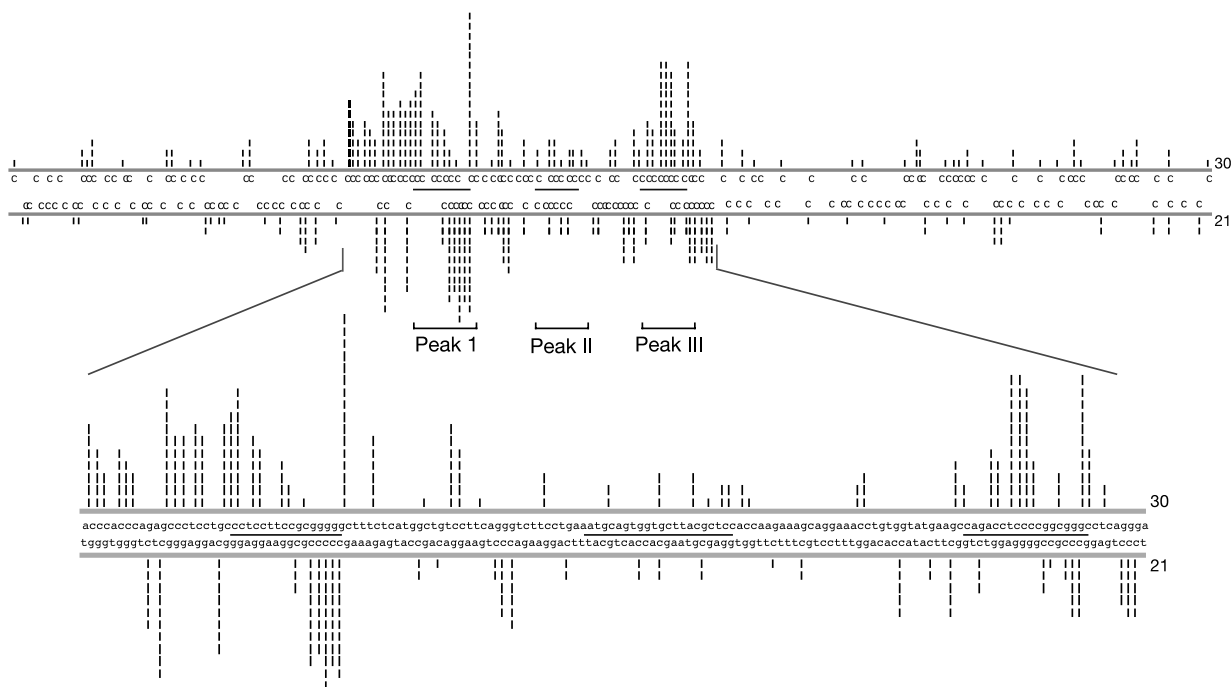


Figure 2 Bisulphite reactivity at the *Bcl-2* Mbr on chromosomal DNA. A 528-bp fragment containing the Mbr amplified from chromosomal DNA after bisulphite treatment is shown. The 150-bp *Bcl-2* Mbr region is expanded to show the complete sequence. The three breakpoint peaks are indicated by three short horizontal lines between the top and bottom strands (see Supplementary Fig. 1). Bisulphite sensitivity is shown for the 250 bp

downstream of the Mbr and 125 bp upstream of the Mbr. Vertical incremental bars (vertical dashes) above the line indicate the sensitivity for the top strand, and bars below the line indicate the sensitivity for the bottom strand; each vertical bar represents a cytosine conversion on one molecule. The total numbers of molecules sequenced from the top and bottom strands are indicated at the right margin.

we find that 28% of human chromosomal alleles have long single-stranded regions at peaks I and III of the *Bcl-2* Mbr. (Overall 15.5% of the cytosines are converted; Fig. 3.) There is also a tendency for consecutive cytosines to be converted (seven or more cytosine residues distributed over 14–49 bp) (Figs 2 and 3). These regions of single-strandedness suggest the existence of a non-B-form single-stranded DNA conformation at the Mbr in this human pre-B-cell line.

Controls at nine unrelated genomic sites showed that the *Bcl-2* Mbr is distinctive for the lengths of DNA that are single-stranded (see Supplementary text and Supplementary Fig. 5). We confirmed the single-stranded character of the *Bcl-2* Mbr in chromatin by diffusing two other types of chemical probes (KMnO_4 and OsO_4) into viable cells. Results from these experiments also indicated that there is a single-stranded character at the Mbr (see Supplementary text and Supplementary Fig. 6).

To further investigate the requirements of non-B-DNA structure formation at the *Bcl-2* Mbr, plasmid pXW5, bearing the *Bcl-2* fragment, was propagated as a replicating human minichromosome in Reh cells, harvested 42 h later and subjected to the bisulphite modification assay. The results are indistinguishable from those observed above for the chromosomal DNA (see Supplementary Fig. 7). This indicates that, on an episome, the 300-bp region

containing the *Bcl-2* Mbr is sufficient to assume the single-stranded conformation that is observed in the human chromosome, regardless of the sequence of the neighbouring regions. Again both B-form and non-B-form conformations exist among the population of molecules and in a similar proportion to that observed in the chromosome (Figs 2 and 3).

We wondered whether plasmid DNA bearing the *Bcl-2* Mbr and harvested from bacteria would be able to form the non-B-DNA structure. pXW5 DNA was extracted using a non-denaturing method and subjected to the bisulphite modification assay. The results obtained after sequencing a 930-bp fragment show a high bisulphite sensitivity in the *Bcl-2* Mbr (Fig. 4a). The single-stranded regions are restricted to peaks I and III (Fig. 4a, c, d), which is the case for the human chromosomal and minichromosomal studies of the same DNA region. The conversion frequency at the *Bcl-2* Mbr is 22.5% when the plasmid DNA is supercoiled, and this is 3.9-fold higher than the background. The Mbr in linearized plasmids is still hyper-reactive with bisulphite relative to the surrounding DNA, although the overall reactivity is reduced by 1.8-fold (Fig. 4b).

PCR fragments of shorter length are able to reproduce the non-B conformation based on the bisulphite reactivity, gel mobility shift and by P1-sensitivity analysis (see Supplementary text).

We have reproduced key aspects of the t(14;18) translocation on

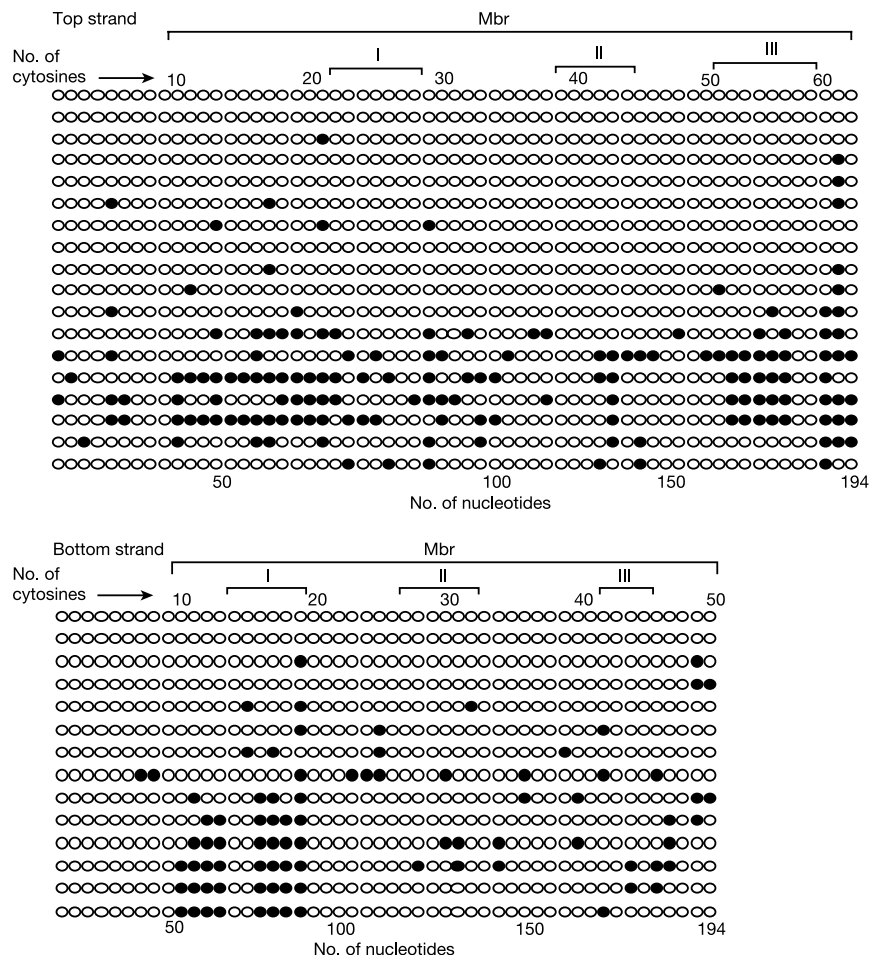


Figure 3 The bimodal nature (B-form DNA versus non-B-form) of the structural configuration of chromosomal DNA at the Mbr. The upper panel shows bisulphite sensitivity of molecules from the top strand, whereas the lower panel shows the corresponding bottom strand data. In both panels, each row of circles represents one DNA molecule. Each filled circle represents an instance of a bisulphite-converted cytosine (single-stranded), whereas open circles represent cytosines resistant to bisulphite (that is,

double-stranded DNA). Note that only the C residues are depicted, but a scale to relate the length along the DNA is shown at the bottom. The designations I, II or III correspond to the three translocation frequency peaks within the Mbr. The population of molecules includes a mixture of those that are B-form and those that have substantial focal and recurring regions of single-strandedness (non-B-form).

extrachromosomal DNA substrates transfected into human lymphoid cells. On substrates bearing the *Bcl-2* Mbr and only one of either the 12- or 23-signal, no *in vivo* recombinants using the 12- or 23-signal were detected, even though breaks at the Mbr were quite focused. However, when we included a pair of 12- and 23-signals, Mbr recombination specific to the 23-signal was readily detectable. The simplest explanation for this is that one or both of the coding ends are released during a 12/23 paired V(D)J recombination event¹³ and simultaneously a break at the *Bcl-2* Mbr is generated by the RAGs. Subsequently, the break at the Mbr is joined to a coding end from the failed V(D)J recombination reaction in ways that are diagrammatically indistinguishable from the t(14;18) translocation (Supplementary Fig. 1a). The Mbr break does not seem to involve direct pairing with any 12- or 23-signal; otherwise, we would have seen recombination on substrates that carry the Mbr and only one signal.

We have noted that the single-strandedness at both peaks I and III varies between individual molecules (Figs 3 and 4). For peak II, there is no marked degree of single-strandedness, and yet it, like peaks I and III, is subject to translocation. We propose that the single strands at peaks I and III interact such that the region between them (peak II) is a site of marked DNA bending. Although the correlation between the single-stranded regions and the translocations is quite good at peaks I and III, there remains the possibility that *in vivo* proteins bound to these single-stranded and adjacent

double-stranded regions may modify further the conformations, thereby explaining the lack of single-strandedness at peak II.

Why would the RAG complex cleave the non-B-form DNA structure at the *Bcl-2* Mbr? There are two known structures that the RAG complex has been demonstrated to cleave in a sequence-independent manner, and both have stable single-stranded character, in common with the structure here. First, the RAG complex (and other transposase enzymes) cleaves single-stranded 3' overhangs in Mg^{2+} -containing buffers¹⁴. This may reflect some common structural features between 3' overhangs and the presumed single-strandedness thought to exist at the border of the heptamer, where RAGs normally cleave¹⁵. Second, the RAG complex opens DNA hairpins in Mn^{2+} -containing buffers¹⁶. This RAG-mediated hairpin opening is inefficient in Mg^{2+} buffers, and it may not be physiologically relevant to hairpin opening¹⁷, but it illustrates that RAG proteins have some low level of activity on such non-B configurations. Could the non-B-form structure at the *Bcl-2* Mbr somehow be a target for nicking? This appears to be the case, given the efficient nicking at or very near the three peaks of the *Bcl-2* Mbr. We note that we cannot rule out the possibility that any of a number of other structure-specific nucleases contribute to cleavage at the Mbr; however, the low level of focused cleavage in cells that do not express RAGs would argue that other nucleases can only play a minor role, if any at all.

In this study, we have clearly demonstrated that non-B-form

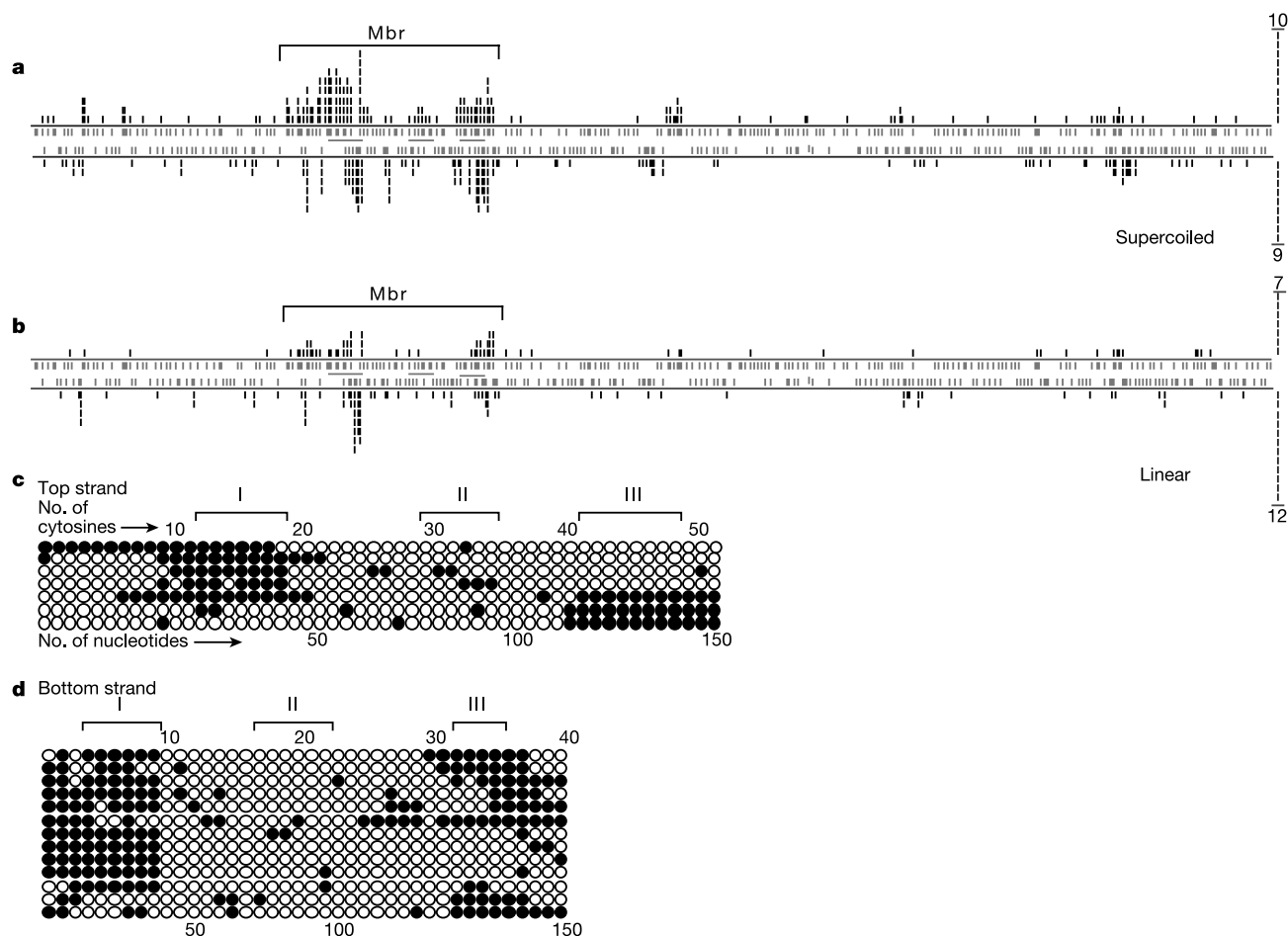


Figure 4 The bisulphite sensitivity of the *Bcl-2* Mbr on plasmid DNA. The pXW5 plasmid DNA was purified from *Escherichia coli* by a non-denaturing method (see Supplementary Methods). Following bisulphite modification, a 930-bp DNA fragment containing the *Bcl-2* Mbr was PCR-amplified and sequenced. **a, b**, Supercoiled plasmid DNA (**a**) and linear

plasmid DNA (**b**) (linearized by *Bgl*II digestion) is shown. **c, d**, Distribution of bisulphite reactivity on molecules with seven or more C to T conversions in a string on the top (**c**) or bottom (**d**) strands of the *Bcl-2* Mbr on supercoiled pXW5. (See Supplementary Fig. 7 legend for more details.)

DNA structural alterations can occur in the human genome and that these alterations can demarcate the precise boundaries of sites of recurrent chromosomal breakage. We have also provided evidence that the RAG complex is responsible for the t(14;18) translocation *in vivo* and *in vitro* and that its role here is different from ones invoked by previously described mechanisms. □

Methods

Plasmid construction and V(D)J recombination assay

The plasmid constructs were made by modifying the SV40-based plasmid, pGG51 (ref. 8). See Supplementary Methods.

Transfection of the 293T cells with pXW5 or pSCR45 along with full-length RAG1/2 (Fig. 1), mutant RAG1/full-length RAG2 (Fig. 1 and Table 1) or core RAG1/2 (Table 1) was done using the calcium-phosphate method as described earlier⁹. The coordinates of the core RAGs are the same as those used for protein production below. Plasmid DNA was alkaline harvested and analysed as described in the Supplementary Methods.

Bisulphite modification assay

The bisulphite modification assay was used as described previously¹¹ (see Supplementary Methods).

Ligation-mediated PCR

For details of the ligation-mediated PCR see Supplementary Methods.

In vitro RAG nicking assay

Core murine glutathione S-transferase (GST)–RAG1 (amino acids 330–1040) and GST–RAG2 (amino acids 1–383), or MBP murine core RAG1 and RAG2, or core RAG1 and full-length MBP RAG2 proteins were overexpressed in the human 293T cells and purified as previously described^{10,18}. See Supplementary Methods.

Received 18 October 2003; accepted 19 January 2004; doi:10.1038/nature02355.

- Lieber, M. R. In *The Causes and Consequences of Chromosomal Aberrations* (ed. Kirsch, I.) 239–275 (CRC Press, Boca Raton, 1993).
- Lewis, S. M. The mechanism of V(D)J joining: lessons from molecular, immunological and comparative analyses. *Adv. Immunol.* **56**, 27–150 (1994).
- Dalla-Favera, R. In *The Causes and Consequences of Chromosomal Aberrations* (ed. Kirsch, I. R.) 313–332 (CRC Press, Boca Raton, 1993).
- Korsmeyer, S. J. Chromosomal translocations in lymphoid malignancies reveal novel proto-oncogenes. *Annu. Rev. Immunol.* **10**, 785–807 (1992).
- Tycko, B. & Sklar, J. Chromosomal translocations in lymphoid neoplasia: A reappraisal of the recombination model. *Cancer Cells* **2**, 1–8 (1990).
- Raghavan, S. C., Kirsch, I. R. & Lieber, M. R. Analysis of the V(D)J recombination efficiency at lymphoid chromosomal translocation breakpoints. *J. Biol. Chem.* **276**, 29126–29133 (2001).
- Jager, U. *et al.* Follicular lymphomas BCL-2/IgH junctions contain templated nucleotide insertions: novel insights into the mechanism of t(14;18) translocation. *Blood* **95**, 3520–3529 (2000).
- Gauss, G. H. & Lieber, M. R. Mechanistic constraints on diversity in human V(D)J recombination. *Mol. Cell. Biol.* **16**, 258–269 (1996).
- Schwarz, K. *et al.* RAG mutations in human B cell-negative SCID. *Science* **274**, 97–99 (1996).
- Swanson, P. C. The DDE motif in RAG-1 is contributed in trans to a single active site that catalyzes the nicking and transesterification steps of V(D)J recombination. *Mol. Cell. Biol.* **21**, 449–458 (2001).
- Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T. E. & Lieber, M. R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nature Immunol.* **4**, 442–451 (2003).
- Gough, G. W., Sullivan, K. M. & Lilley, D. M. The structure of cruciforms in supercoiled DNA: probing the single-stranded character of nucleotide bases with bisulphite. *EMBO J.* **5**, 191–196 (1986).
- Tevelev, A. & Schatz, D. G. Intermolecular V(D)J recombination. *J. Biol. Chem.* **275**, 8341–8348 (2000).
- Santagata, S. *et al.* The RAG1/RAG2 complex constitutes a 3' flap endonuclease: implications for junctional diversity in V(D)J and transpositional recombination. *Mol. Cell* **4**, 935–947 (1999).
- Ramsden, D. A., McBlane, J. F., van Gent, D. C. & Gellert, M. Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. *EMBO J.* **15**, 3197–3206 (1996).
- Besmer, E. *et al.* Hairpin coding end opening is mediated by the recombination activating genes RAG1 and RAG2. *Mol. Cell* **2**, 817–828 (1998).
- Ma, Y., Pannicke, U., Schwarz, K. & Lieber, M. R. Hairpin opening and overhang processing by an Artemis:DNA-PKcs complex in V(D)J recombination and in nonhomologous end joining. *Cell* **108**, 781–794 (2002).
- Yu, K. & Lieber, M. R. The nicking step of V(D)J recombination is independent of synapsis: implications for the immune repertoire. *Mol. Cell. Biol.* **20**, 7914–7921 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. J. Korsmeyer for guidance during the early phases of this work. We also thank I. Haworth, J. S. Lee, P. Chastian and D. Shibata for discussions of the work, and NIH for grants to M.R.L.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.R.L. (lieber@usc.edu).

Tension between two kinetochores suffices for their bi-orientation on the mitotic spindle

Hilary Dewar¹, Kozo Tanaka¹, Kim Nasmyth² & Tomoyuki U. Tanaka¹

¹School of Life Sciences, University of Dundee, Wellcome Trust Biocentre, Dundee DD1 5EH, UK

²Research Institute of Molecular Pathology, Dr Bohr-Gasse 7, A-1030 Vienna, Austria

The movement of sister chromatids to opposite spindle poles during anaphase depends on the prior capture of sister kinetochores by microtubules with opposing orientations (amphitelic attachment or bi-orientation)¹. In addition to proteins necessary for the kinetochore–microtubule attachment, bi-orientation requires the Ipl1 (Aurora B in animal cells) protein kinase^{2–7} and tethering of sister chromatids by cohesin^{8,9}. Syntelic attachments, in which sister kinetochores attach to microtubules with the same orientation, must be either ‘avoided’ or ‘corrected’. Avoidance might be facilitated by the juxtaposition of sister kinetochores such that they face in opposite directions; kinetochore geometry is therefore deemed important. Error correction, by contrast, is thought to stem from the stabilization of kinetochore–spindle pole connections by tension in microtubules, kinetochores, or the surrounding chromatin arising from amphitelic but not syntelic attachment^{10,11}. The tension model predicts that any type of connection between two kinetochores suffices for efficient bi-orientation. Here we show that the two kinetochores of engineered, unreplicated dicentric chromosomes in *Saccharomyces cerevisiae* bi-orient efficiently, implying that sister kinetochore geometry is dispensable for bi-orientation. We also show that Ipl1 facilitates bi-orientation by promoting the turnover of kinetochore–spindle pole connections in a tension-dependent manner.

Evidence for an error correction mechanism facilitating bi-orientation has been hitherto confined to meiotic cells, where homologous kinetochores are connected by chiasmata, whose remoteness from the sites of microtubule attachment precludes a key role for kinetochore geometry. Thus, in grasshopper spermatocytes whose kinetochores and spindle poles are repeatedly connected and disconnected by microtubules until they bi-orient, the kinetochore-to-pole connections of a ‘maternal’ chromosome can be stabilized by using a glass needle to pull on the ‘paternal’ chromosome attached to it^{10,11}. If bi-orientation were established in mitotic cells by a similar tension-sensitive error correction mechanism, then any kind of connection between two kinetochores would suffice for their bi-orientation, including two kinetochores situated on the same chromatid. We therefore set out to analyse the behaviour of dicentric minichromosomes that are unable to replicate during S phase.

We addressed this issue in the budding yeast *S. cerevisiae*, whose kinetochores possess only a single microtubule-binding site¹² and whose centromeres (the DNA underlying the kinetochore) have been pinpointed to sequences no longer than 120 base pairs (bp)¹³. Their activity can therefore be turned off by transcription from an adjacent promoter¹⁴. The behaviour of yeast centromeres can be followed by marking them with bacterial operators bound by repressor proteins fused to green fluorescent protein (GFP)^{8,15,16}. Traction caused by ‘amphitelic’ attachment overwhelms sister chromatid cohesion at centromeres, but not in the flanking sequences, which leads to their precocious separation before onset of anaphase. The tension so created causes centromeric chromatin to unravel, forming 10-nm (beaded nucleosomes) or even thinner fibres.

To address the behaviour of unreplicated dicentric chromosomes, we created a circular minichromosome carrying two centromeres, one constitutive and the other conditionally inactivated owing to its juxtaposition to the *GAL1-10* promoter (ref. 14 and Fig. 1a). The chromosome also possessed a single replication origin flanked by recombination sites for the *Zygosaccharomyces rouxii* R recombinase¹⁷ and contained a tandem array of *tet* operators that bind Tet repressor–GFP fusion proteins¹⁸. The host cell expressed the recombinase gene from the *MET3* promoter, which is transcribed only in the absence of methionine. The minichromosome was stably propagated as long as cells were grown in the presence of galactose, which inactivates one of the centromeres, and in the presence of methionine, which prevents expression of the recombinase.

Removal of methionine from the culture medium caused efficient recombination between recombination sites and the accumulation of minichromosomes lacking any replication origin (Fig. 1b, 2 × RS), leading to a rapid increase in cells born either without any minichromosome or with an unreplicated minichromosome (see below). The origin was not lost from minichromosomes containing only a single recombination site (Fig. 1b, 1 × RS). After most minichromosomes had lost their origins, galactose was replaced by glucose in the medium, which rapidly activated the second centromere of the minichromosome (Supplementary Note 1 and Fig. S1).

The behaviour of unreplicated dicentric (2 × CEN) and mono-

centric (1 × CEN) minichromosomes was observed by time-lapse microscopy. Both types of minichromosome were marked by the Tet repressor fused to GFP, and spindle pole bodies (SPBs) were marked by Spc42 fused to yellow fluorescent protein (YFP). During metaphase, the GFP signals that were associated with all 38 unreplicated monocentric minichromosomes remained in the close vicinity of one SPB (Fig. 2a, top). Although they changed their SPB partner on rare occasions (see Supplementary Note 5), they never spent any appreciable time in the space between the SPBs.

The GFP signals of all 34 unreplicated dicentric minichromosomes, by contrast, spent most of this period halfway between the two SPBs (Fig. 2a, bottom). They frequently moved vigorously back and forth along the axis connecting the SPBs, and most (75%) were stretched along this axis. The compaction¹⁹ of stretched operators was reduced between 6- and 30-fold, and was reduced even further in *yku80* deletion mutant cells (data not shown), as

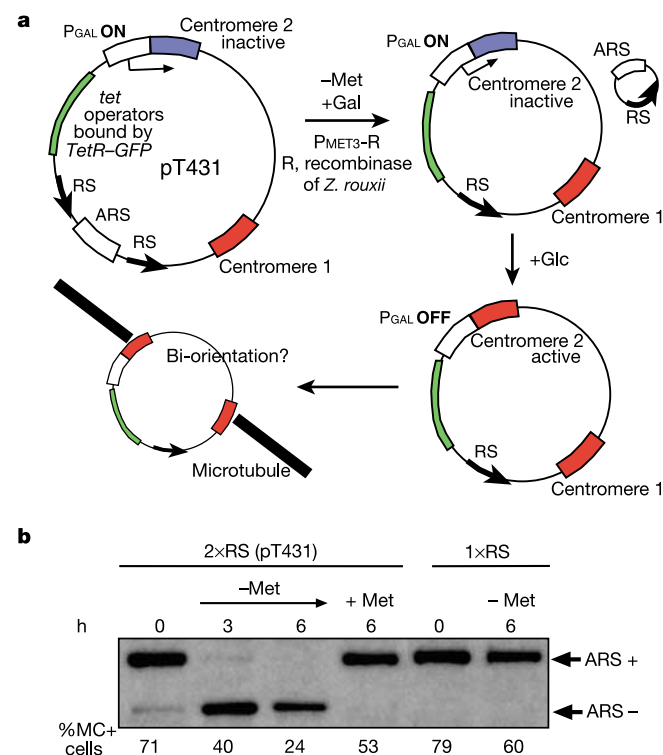


Figure 1 Making unreplicated dicentric minichromosomes. **a**, Scheme for making an unreplicated dicentric minichromosome (see text). After removing the DNA replication origin, the intervals between centromeres are 7.0 kb on the side containing *tet* operators and 4.2 kb on the other side. ARS, DNA replication origin; RS, recombination site; *P_{MET3}*, *MET3* promoter; *P_{GAL}*, *GAL1-10* promoter; Met, methionine; Gal, galactose; Glc, glucose. **b**, Southern blot analysis and percentage of cells containing minichromosomes. *TetR-GFP P_{MET3}-R* cells carrying pT431 or pT432 were incubated in medium with galactose either with (+Met) or without (–Met) methionine. The percentage of cells containing minichromosomes out of all cells observed (% MC+ cells) is shown for each time point. 2 × RS (pT431), DNA replication origin flanked by two recombination sites (see **a**); 1 × RS (pT432), DNA replication origin flanked by only one recombination site.

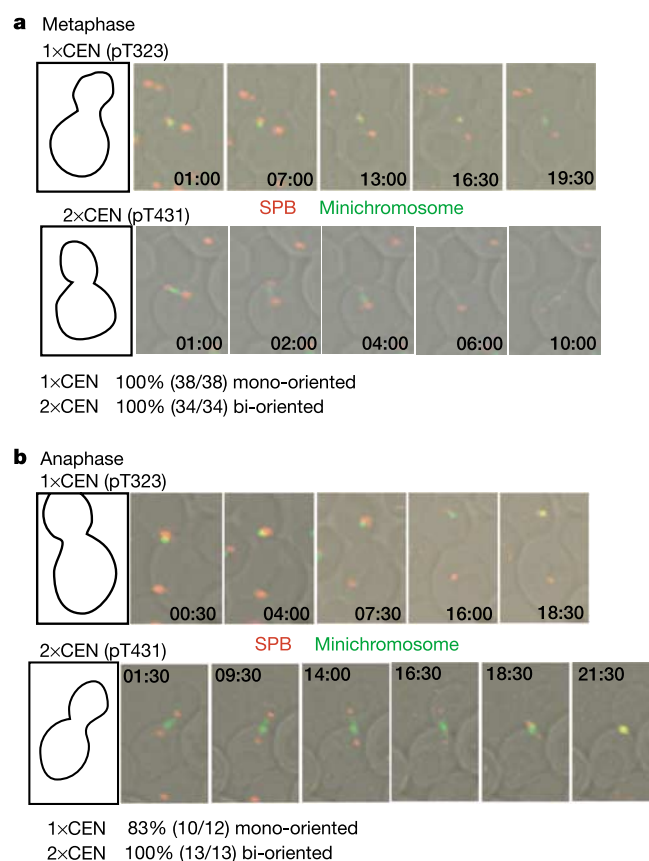


Figure 2 Behaviour of unreplicated monocentric and dicentric minichromosomes. **a**, Behaviour in metaphase. *TetR-GFP SPC42-YFP P_{MET3}-R* cells carrying pT323 (T2755, top) or pT431 (T3000, bottom) were incubated in the absence of methionine, first with galactose for 4 h and then with glucose. Time-lapse images were collected for 20 min at 30-s intervals, commencing 20 min after transfer to glucose medium. Shown are representative time-lapse images of cells with unreplicated minichromosomes. The time in minutes from the start of image acquisition is indicated. When both separated SPBs were in the mother cell, the cell was considered to be in metaphase. The behaviour of minichromosomes (mono-oriented or bi-oriented) was scored as described in Methods. When minichromosome behaviours did not belong to either category (less than 3% of total counts) or when the bipolar spindle was short (less than 1 μm) on a projected image, they were not scored. See also Supplementary Movie 2a. **b**, Behaviour in anaphase. T2755 and T3000 cells were treated as in **a**. When spindles were at least twice as long as those in metaphase at least at one time point, with one SPB in the bud and the other in the mother cell, the cell was considered to have initiated anaphase. The behaviour of unreplicated minichromosomes was scored as in **a**. See also Supplementary Movie 2b.

previously reported for replicated dicentric chromosomes²⁰. These behaviours of GFP signals emanating from unreplicated dicentric minichromosomes imply that their two kinetochores are simultaneously attached to opposite poles and pulled in opposite directions; in other words, they bi-orient on the spindle.

During anaphase, 10 out of 12 monocentric unreplicated minichromosomes remained in the vicinity of one of the two SPBs. These monocentric minichromosomes followed the SPB as it moved towards the cell cortex (Fig. 2b, top). For unknown reasons,

2 out of 12 monocentric minichromosomes did not segregate with an SPB. By contrast, all 13 dicentric unreplicated minichromosomes remained in the space between the SPBs on initiation of anaphase (Fig. 2b, bottom). Their GFP signals usually remained stretched and sometimes moved vigorously between SPBs. In most cells containing these dicentric minichromosomes, the SPBs started to segregate from each other but their movement was halted before they reached the cell cortex (Supplementary Note 2).

Do unreplicated dicentric minichromosomes establish bi-orientation as efficiently as replicated monocentric ones in cell cycles? To answer this question, we compared the kinetics of their bi-orientation establishment in synchronized cells. We arrested cells containing unreplicated monocentric or dicentric minichromosomes or replicated monocentric minichromosomes with α -factor and released them into fresh medium. The rate of bi-oriented minichromosomes was scored during the bud emergence and establishment of a bipolar spindle. As expected, unreplicated monocentric minichromosomes did not show any appreciable bi-orientation (Fig. 3a). Both unreplicated dicentric and replicated monocentric minichromosomes established bi-orientation soon after bipolar spindle formation (SPB separation) with similar kinetics (Fig. 3b, c). We also found that many cells underwent SPB separation during time-lapse image acquisition. Most if not all unreplicated dicentric (24 out of 25 observed; Supplementary Fig. S2) and replicated monocentric (14 out of 15 observed) minichromosomes showed evidence of bi-orientation (see Methods) within 2 min of SPB

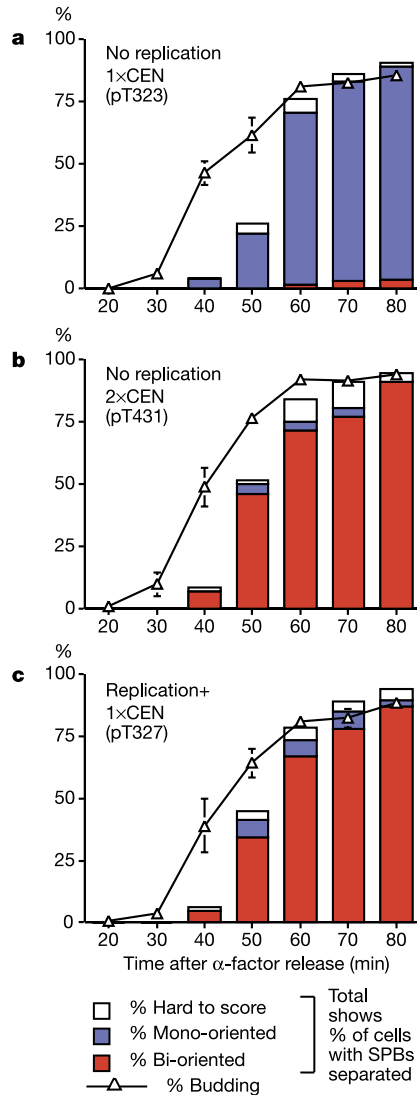


Figure 3 Comparing the kinetics of bi-orientation establishment during a synchronous cell cycle. *TetR-GFP SPC42-YFP P_{MET3}-R* cells carrying pT323 (T2755; **a**), pT431 (T3000; **b**) or pT327 (T2951; **c**) were incubated in the absence of methionine with galactose for 2 h and subsequently treated with α -factor for 2.5 h. They were then released into fresh YPA medium containing glucose. After the release, samples were collected for bud counts and time-lapse image acquisition every 10 min. Commencing immediately after sample collection, live cell images were acquired at 30-s intervals for 6 min. The behaviour of the three minichromosomes was scored as described in Methods. In addition, pre-anaphase separation of sister minichromosomes³¹ was counted as bi-orientation (only for pT327). From three trials of the release from α -factor arrest of each strain, 50–80 minichromosomes were scored at each time point, including ones with unseparated SPBs. Scoring was difficult if the distance between SPBs was too short in the projected images ('hard to score'). If SPBs separated within 2 min of the start of image acquisition, the cells were considered to have separated SPBs. The mean $\pm$ s.d. percentage of cells with buds from three trials of α -factor release is plotted at each time point.

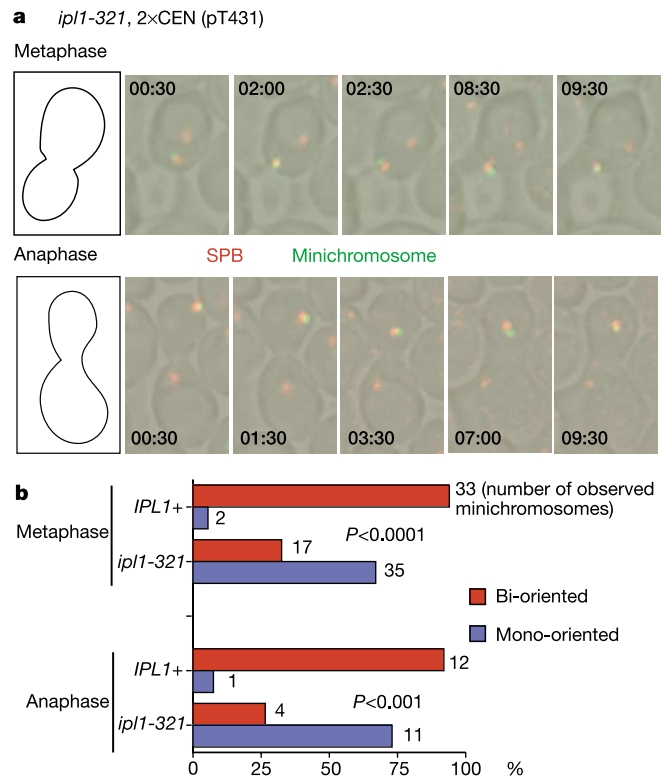


Figure 4 Behaviour of unreplicated dicentric minichromosomes in *ip11* mutant cells. *IP11+* (T3000) and *ip11-321* (T3001) cells (*TetR-GFP SPC42-YFP P_{MET3}-R* pT431) were incubated in medium without methionine but with galactose at 25 °C for 3 h and then at 35 °C (the restrictive temperature for *ip11-321*) for 1 h, and subsequently with glucose at 35 °C. The images were collected at 35 °C, commencing 20 min after transfer to glucose medium. The behaviour of minichromosomes was scored as in Fig. 2. **a**, Representative time-lapse images. **b**, Scoring. When synchronous cultures (release from α -factor arrest) of the two strains were observed in glucose medium at 35 °C, similar results were obtained in metaphase and anaphase during the first cell cycle (data not shown).

separation. These data suggest that unreplicated dicentric minichromosomes establish bi-orientation as efficiently as replicated monocentric minichromosomes.

It could be argued that proximity of the two centromeres on the unreplicated dicentric minichromosomes precludes their syntelic attachment and that this ensures efficient bi-orientation. The shortest distance between them is 4.2 kilobases (kb), which might be expected to span 12–18 nm (ref. 19). Because this is smaller than the diameter of microtubules (25 nm), there might be insufficient space for both centromeres to attach to microtubules with the same orientation. To address this issue, we enlarged the distance between centromeres to 10 kb (28–42 nm) on both sides of the circular minichromosome, but this had no effect on the efficiency of bi-orientation (27 out of 27 bi-oriented). It is therefore unlikely that efficient bi-orientation is caused by centromere proximity. Instead, our data imply that the connection of two kinetochores alone is sufficient for their efficient bi-orientation. The simplest explanation for this phenomenon is that any connection between kinetochores can provide the tension needed to stabilize kinetochore-to-pole connections.

This tension is normally made possible by connections between sister chromatids that are mediated by cohesin, whose absence causes frequent mono-orientation of replicated kinetochores^{8,9}. If the primary role of cohesin in promoting bi-orientation is to supply tension, then it should be possible to restore bi-orientation to cohesin-depleted cells by providing an alternative means of connection between sister DNAs. A possible way to achieve this would be by inactivating topoisomerase II, which is required to decatenate sister chromatids after DNA replication^{21–23}. We therefore compared the efficiency of sister kinetochore bi-orientation in *TOP2+* and *top2-4* (ref. 22) cells after depleting the Scc1 subunit of cohesin (Supplementary Note 3 and Fig. S3). The centromeres on chromosome V (*CEN5s*) were bi-oriented in 55% of *TOP2+* cells and in 77% of *top2-4* cells. Thus bi-orientation was partially restored when topoisomerase II function was impaired in cohesin-depleted cells. A similar result was obtained when Scc1-depleted DT40 chicken cells were treated with an inhibitor of topoisomerase II (ref. 24). These data suggest that any kind of connection between sister kinetochores may be sufficient to induce their bi-orientation.

To address whether the bi-orientation of unreplicated dicentric minichromosomes is, like that of authentic replicated sister kinetochores^{2,3,6}, dependent on the Ipl1 kinase, we compared the behaviour of unreplicated dicentric minichromosomes in *IPL1+* and *ipl1-321* mutant² cells. Bi-orientation of unreplicated dicentric minichromosomes took place in 94% of *IPL1+* cells (45 out of 48 in metaphase and anaphase; Fig. 4b). This occurred in only 31% of *ipl1-321* cells (21 out of 67). In the remaining 69%, minichromosome GFP signals remained in the vicinity of one SPB (Fig. 4a, b). They made connections to a single pole; that is, they mono-oriented (Supplementary Note 4). Inactivation of Ipl1 therefore greatly reduces the incidence with which unreplicated dicentric minichromosomes bi-orient. Ipl1 seems to be just as important for the bi-orientation of unreplicated dicentric minichromosomes as it is for replicated monocentric minichromosomes. We obtained similar results with a mutant of Sli15 (data not shown), which is an orthologue of Inner centromere protein (INCENP) and forms a complex with Ipl1 (refs 25–27). The Ipl1–Sli15 complex must be therefore capable of promoting bi-orientation through mechanisms that do not involve sister kinetochore geometry.

It has been suggested that Ipl1 promotes bi-orientation by facilitating the turnover of kinetochore–spindle pole connections when tension has not been generated⁶, as occurs during syntelic attachment. However, previous work has never properly documented the role of Ipl1 in detaching a kinetochore from one SPB and attaching it to the opposite pole after SPB separation. This process should occur with unreplicated monocentric minichromosomes on

which tension cannot be exerted. We developed an assay by which we could directly visualize the process of re-orientation that unreplicated monocentric minichromosomes undergo after SPB separation (Supplementary Note 5 and Fig. S4). We found that *ipl1-321* mutants showed less frequent re-orientation than *IPL1+* cells. These data suggest that Ipl1 does indeed have an important role in promoting the turnover of kinetochore–spindle pole connections when they cannot come under tension normally generated by amphitelic attachment.

Three decades ago it was shown that the connection of a kinetochore to a spindle pole by microtubules during meiosis I of grasshopper spermatocytes is stabilized by tension¹⁰. This implied that error correction has a key role in bi-orienting maternal and paternal chromosomes during meiosis I. It has long been disputed whether a similar principle operates during mitosis¹. Because sister kinetochores lie back to back, it has been thought that kinetochore geometry and not error correction might be the driving force behind the bi-orientation of sister kinetochores. Studies have implicated the Ipl1 (Aurora B) kinase and cohesin in promoting mitotic bi-orientation^{2–9}. Do they act by creating a back-to-back kinetochore geometry conducive to bi-orientation, or instead as part of an error correction mechanism? The finding that Ipl1 promotes the realignment of kinetochore–spindle pole connections in budding yeast, from old to new spindle poles, is consistent with an error correction mechanism⁶. Our results provide further evidence that erroneous connections are indeed corrected in a tension-dependent manner.

Our data suggest that the main, if not only, role of cohesin during mitosis in *S. cerevisiae* is to provide the physical connection between sister kinetochores necessary to generate tension when they bi-orient, and that an important function of Ipl1 is to eliminate kinetochore–spindle pole connections that do not generate such tension. This notion is also supported by the report that syntelic attachments are removed by Aurora-B-dependent mechanisms in mammalian cells⁷. How tension is actually sensed and how this affects whether or not Ipl1 promotes re-orientation remain to be addressed. Phosphorylation of the Dam1 kinetochore protein could be one of the key events in this process²⁸. Our data do not of course exclude the possibility that sister kinetochore geometry is also conducive to bi-orientation. For instance, Ipl1 (Aurora B) and/or cohesin might have additional roles in promoting kinetochore geometry²⁹. In particular, fission yeast and animal cells must have a single kinetochore attached to more than one microtubule. Chromosome bi-orientation fails if a single kinetochore is captured by microtubules from the opposite spindle poles (merotelic attachment)³⁰. In these organisms, sister kinetochore geometry may have more important roles to ensure bi-orientation than in *S. cerevisiae* cells (see Supplementary Note 6). □

Methods

Yeast genetics and molecular biology

The background of yeast strains (W303), yeast media, Southern blotting, α -factor arrest/release, *TetR*–*GFP*/*tet* operator system, and the construction of pT323, pT327, *CEN5*–*tetOs* and *SPC42*–*YFP* have been described^{6,8,16,18}. To construct *P_{MET3}*–*R*, pT343 was made by cloning the *R* recombinase gene (from pHM153; ref. 17) under control of the *MET3* promoter into YIplac204 (National Center for Biotechnology Information, X75461) and integrated at the *trp1* locus. pT431 and pT432 were constructed by cloning *CEN4* (730-bp DNA containing *CDE1-II-III*) under control of the *GAL1*–10 promoter into pT323 and pT327, respectively. pT323 and pT327 are the same, except that the former has two recombination sites that flank a replication origin, whereas the latter has only one⁶. See Supplementary Note 7 for further methodological details.

Microscopy

The general procedures for time-lapse microscopy have been described⁸. We collected time-lapse images every 20–30 s, either for 20–30 min at 23 °C (ambient temperature) or for 10–15 min at 35 °C or 37 °C unless otherwise stated. One bright field image and 7–9 z-stacks (each 0.5–0.7- μ m apart) of GFP or YFP images were acquired with a J3 filter set (Chroma) at each time point. The z-stacks were deconvoluted and projected to two-dimensional images⁸ with the Openlab (Improvision) or SoftWoRx (Applied Precision) software.

Minichromosome scoring

The behaviour of minichromosomes (mono-oriented or bi-oriented) was scored as follows unless otherwise stated. When minichromosomes were not more than one-fifth of the spindle length away from an SPB at any two consecutive time points, they were scored as 'mono-oriented'. When minichromosomes were more than one-fifth of the spindle length away from both SPBs at most time points or they moved vigorously between two SPBs, they were scored as 'bi-oriented'. To score the behaviour of pT323 and pT431, rare replicated minichromosomes were distinguished (and ignored) owing to the greater intensity of their GFP signals. For pT327, a few cells with excessive copies of minichromosomes were not included in scoring.

Received 14 October 2003; accepted 12 January 2004; doi:10.1038/nature02328.

Published online 11 February 2004.

1. Tanaka, T. U. Bi-orienting chromosomes on the mitotic spindle. *Curr. Opin. Cell Biol.* **14**, 365–371 (2002).
2. Biggins, S. *et al.* The conserved protein kinase Ipl1 regulates microtubule binding to kinetochores in budding yeast. *Genes Dev.* **13**, 532–544 (1999).
3. He, X., Rines, D. R., Espelin, C. W. & Sorger, P. K. Molecular analysis of kinetochore–microtubule attachment in budding yeast. *Cell* **106**, 195–206 (2001).
4. Adams, R. R., Maiato, H., Earnshaw, W. C. & Carmena, M. Essential roles of *Drosophila* inner centromere protein (INCENP) and aurora B in histone H3 phosphorylation, metaphase chromosome alignment, kinetochore disjunction, and chromosome segregation. *J. Cell Biol.* **153**, 865–880 (2001).
5. Kaitna, S., Pasierbek, P., Jantsch, M., Loidl, J. & Glotzer, M. The aurora B kinase AIR-2 regulates kinetochores during mitosis and is required for separation of homologous chromosomes during meiosis. *Curr. Biol.* **12**, 798–812 (2002).
6. Tanaka, T. U. *et al.* Evidence that the Ipl1–Slh15 (Aurora kinase–INCENP) complex promotes chromosome bi-orientation by altering kinetochore–spindle pole connections. *Cell* **108**, 317–329 (2002).
7. Hauf, S. *et al.* The small molecule Hesperadin reveals a role for Aurora B in correcting kinetochore–microtubule attachment and in maintaining the spindle assembly checkpoint. *J. Cell Biol.* **161**, 281–294 (2003).
8. Tanaka, T., Fuchs, J., Loidl, J. & Nasmyth, K. Cohesin ensures bipolar attachment of microtubules to sister centromeres and resists their precocious separation. *Nature Cell Biol.* **2**, 492–499 (2000).
9. Sonoda, E. *et al.* Scc1/Rad21/Mcd1 is required for sister chromatid cohesion and kinetochore function in vertebrate cells. *Dev. Cell* **1**, 759–770 (2001).
10. Nicklas, R. B. & Koch, C. A. Chromosome micromanipulation. III. Spindle fiber tension and the reorientation of mal-oriented chromosomes. *J. Cell Biol.* **43**, 40–50 (1969).
11. Nicklas, R. B. How cells get the right chromosomes. *Science* **275**, 632–637 (1997).
12. Winey, M. *et al.* Three-dimensional ultrastructural analysis of the *Saccharomyces cerevisiae* mitotic spindle. *J. Cell Biol.* **129**, 1601–1615 (1995).
13. Hegemann, J. H. & Fleig, U. N. The centromere of budding yeast. *BioEssays* **15**, 451–460 (1993).
14. Hill, A. & Bloom, K. Genetic manipulation of centromere function. *Mol. Cell. Biol.* **7**, 2397–2405 (1987).
15. Goshima, G. & Yanagida, M. Establishing biorientation occurs with precocious separation of the sister kinetochores, but not the arms, in the early spindle of budding yeast. *Cell* **100**, 619–633 (2000).
16. He, X., Asthana, S. & Sorger, P. K. Transient sister chromatid separation and elastic deformation of chromosomes during mitosis in budding yeast. *Cell* **101**, 763–775 (2000).
17. Matsuzaki, H., Nakajima, R., Nishiyama, J., Araki, H. & Oshima, Y. Chromosome engineering in *Saccharomyces cerevisiae* by using a site-specific recombination system of a yeast plasmid. *J. Bacteriol.* **172**, 610–618 (1990).
18. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell* **91**, 35–45 (1997).
19. Guacci, V., Hogan, E. & Koshland, D. Chromosome condensation and sister chromatid pairing in budding yeast. *J. Cell Biol.* **125**, 517–530 (1994).
20. Thrower, D. A. & Bloom, K. Divalent chromosome stretching during anaphase reveals roles of Sir2/Ku in chromatin compaction in budding yeast. *Mol. Biol. Cell* **12**, 2800–2812 (2001).
21. DiNardo, S., Voelkel, K. & Sternglanz, R. DNA topoisomerase II mutant of *Saccharomyces cerevisiae*: topoisomerase II is required for segregation of daughter molecules at the termination of DNA replication. *Proc. Natl Acad. Sci. USA* **81**, 2616–2620 (1984).
22. Holm, C., Goto, T., Wang, J. C. & Botstein, D. DNA topoisomerase II is required at the time of mitosis in yeast. *Cell* **41**, 553–563 (1985).
23. Uemura, T. *et al.* DNA topoisomerase II is required for condensation and separation of mitotic chromosomes in *S. pombe*. *Cell* **50**, 917–925 (1987).
24. Vagnarelli, P. *et al.* Analysis of Scc1-deficient cells defines a key metaphase role of vertebrate cohesin in linking sister kinetochores. *EMBO Rep.* **5**, 167–171 (2004).
25. Kim, J. H., Kang, J. S. & Chan, C. S. Slh15 associates with the Ipl1 protein kinase to promote proper chromosome segregation in *Saccharomyces cerevisiae*. *J. Cell Biol.* **145**, 1381–1394 (1999).
26. Kaitna, S., Mendoza, M., Jantsch-Plunger, V. & Glotzer, M. INCENP and an Aurora-like kinase form a complex essential for chromosome segregation and efficient completion of cytokinesis. *Curr. Biol.* **10**, 1172–1181 (2000).
27. Adams, R. R. *et al.* INCENP binds the Aurora-related kinase AIRK2 and is required to target it to chromosomes, the central spindle and cleavage furrow. *Curr. Biol.* **10**, 1075–1078 (2000).
28. Cheeseman, I. M. *et al.* Phospho-regulation of kinetochore–microtubule attachments by the Aurora kinase Ipl1p. *Cell* **111**, 163–172 (2002).
29. Losada, A., Hirano, M. & Hirano, T. Cohesin release is required for sister chromatid resolution, but not for condensin-mediated compaction, at the onset of mitosis. *Genes Dev.* **16**, 3004–3016 (2002).
30. Cimini, D. *et al.* Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells. *J. Cell Biol.* **153**, 517–527 (2001).
31. Tanaka, T., Cosma, M. P., Wirth, K. & Nasmyth, K. Identification of cohesin association sites at centromeres and along chromosome arms. *Cell* **98**, 847–858 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Toth for discussions leading to the use of unreplicated dicentric chromosomes; W. Earnshaw for discussing inhibition of Top2 in Scc1-depleted cells and for sharing unpublished data; M. J. R. Stark and J. Swedlow for discussion and critically reading the manuscript; N. Mukae and C. Newlon for discussion; P. Andrews and S. Swift for help with time-lapse microscopy; and R. Ciosk, S. Biggins, F. Uhlmann, X. He, P. Sorger, R. Tsien, W. Fangman, H. Araki, C. Holm, R. Sternglanz, C. Chan and EUROSCARF for reagents. This work was supported by The Wellcome Trust, Cancer Research UK, EMBO Young Investigator Program, and a fellowship (to K.T.) from Japan Society for the Promotion of Science.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.U.T. (t.tanaka@dundee.ac.uk).

Preferential *cis-syn* thymine dimer bypass by DNA polymerase η occurs with biased fidelity

Scott D. McCulloch¹, Robert J. Kokoska¹, Chikahide Masutani², Shigenori Iwai³, Fumio Hanaoka^{2,4} & Thomas A. Kunkel¹

¹Laboratory of Molecular Genetics and Laboratory of Structural Biology, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina 27709, USA

²Graduate School of Frontier Biosciences, Osaka University and CREST, Japan Science and Technology Corporation, 1-3 Yamada-oka, Suita, Osaka 565-0871, Japan

³Division of Chemistry, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

⁴Discovery Research Institute, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

Human DNA polymerase η (Pol η) modulates susceptibility to skin cancer by promoting DNA synthesis past sunlight-induced cyclobutane pyrimidine dimers that escape nucleotide excision repair (NER)^{1,2}. Here we have determined the efficiency and fidelity of dimer bypass. We show that Pol η copies thymine dimers and the flanking bases with higher processivity than it copies undamaged DNA, and then switches to less processive synthesis. This ability of Pol η to sense the dimer location as synthesis proceeds may facilitate polymerase switching before and after lesion bypass. Pol η bypasses a dimer with low fidelity and with higher error rates at the 3' thymine than at the 5' thymine. A similar bias is seen with *Sulfolobus solfataricus* DNA polymerase 4, which forms a Watson–Crick base pair at the 3' thymine of a dimer but a Hoogsteen base pair at the 5' thymine (ref. 3). Ultraviolet-induced mutagenesis is also higher at the 3' base of dipyrimidine sequences^{4–6}. Thus, in normal people and particularly in individuals with NER-defective xeroderma pigmentosum who accumulate dimers, errors made by Pol η during dimer bypass could contribute to mutagenesis and skin cancer.

To compare the efficiency with which human Pol η bypasses a *cis-syn* thymine–thymine (TT) dimer or undamaged thymines in the same sequence context, we used a large excess of primer template and incubation times short enough to ensure that the DNA products result from a single cycle of synthesis⁷. Thus, the dimer is encountered while the polymerase is processively synthesizing DNA using all four dNTPs. This provides bypass efficiency parameters that differ from those in previous studies that measured the rate of single nucleotide insertion.

Human Pol η has low processivity with undamaged DNA (ref. 8 and Fig. 1a, left, and 1b–d, open bars). With damaged DNA, processivity is unexpectedly higher at the 3' and 5' T of dimers

and the flanking template bases (Fig. 1a–d). A structural basis for this pattern is suggested by the crystal structure of a DNA decamer containing a TT dimer⁹, which shows abnormal base-pair step and helical parameters at the dimer and flanking base pairs (Fig. 1b–d, red lines). After incorporating one (Fig. 1d) or two nucleotides beyond the 5' T (Fig. 1b, c), Pol η switches to less processive synthesis. The switch is seen in three sequence contexts (Fig. 1b, c), it is dimer dependent (Fig. 1a–e), and it is independent of the location of the primer used to initiate translesion synthesis (TLS; Fig. 1e).

When DNA product distributions are used to calculate insertion efficiency⁷, the values are higher with the dimer-containing substrate for several consecutive insertions (Fig. 1f). The results at the 5' T are consistent with a slightly higher rate of dATP incorporation by yeast Pol η opposite the 5' T of a dimer as compared with the corresponding undamaged thymine in reactions with a single nucleotide¹⁰. The differences are small but reproducible in several determinations, and cumulatively result in relative dimer bypass efficiencies (Fig. 1f) that are 300% (Fig. 1a, b), 160% (Fig. 1c) and 130% (Fig. 1d) of that seen with undamaged thymines. This contrasts with much less efficient relative bypass of a 6-4 photo-product ($\geq 2\%$ relative bypass efficiency, data not shown) or apurinic and/or apyrimidinic sites (10–13%; ref. 7). The results with human Pol η also contrast with values for *S. solfataricus* DNA polymerase 4 (Dpo4), a Y family member that is more processive than Pol η and that can also bypass TT dimers (ref. 11 and Fig. 2a). A dimer clearly impedes synthesis by Dpo4, as judged by the strong, dimer-dependent increases in termination probability at the 3' T and the preceding template base (Fig. 2a, b) and by a relative dimer bypass efficiency of 24% (Fig. 2c), which is 12-fold lower than the value for Pol η (300%, Fig. 1f).

These data imply that polymerase switching during translesion

replication may occur during transitions from preferential to disfavoured use of a damaged template primer, and that the polymerase called on to continue synthesis is one that has evolved to use that particular substrate most efficiently. For example, if Pol δ terminates synthesis before the 3' T of the dimer (Fig. 1g), the fact that the product of that reaction is a preferred substrate for Pol η may facilitate a switch to Pol η rather than to any of the other four human polymerases implicated in TLS (Pol κ , Pol ι and Pol ζ and REV1p; reviewed in ref. 12). Once Pol η copies the dimer and one or two more nucleotides, its ability to sense the dimer present in the nascent duplex results in an increase in termination that may facilitate the switch to Pol δ or another polymerase.

Termination at +1 and +2 (Fig. 1f) corresponds to damaged bases at the second, third and fourth base pairs of the duplex primer template—positions that would interact with the little finger (also called the polymerase associated domain (PAD)) domain^{13,14} unique to Y family polymerases. These interactions could trigger a switch to another polymerase, perhaps through conformational changes, rather than a switch dependent on stochastic enzyme dissociation. This idea may be relevant to other TLS polymerase and lesion combinations, with differences in little-finger domains¹³ contributing to differences in bypass specificity. The exact locations of switches before and after TLS may vary depending on the identity of the competing polymerases and lesions and on the DNA sequence context, as implied by the results in Fig. 1b, c (switch at +2) compared with those in Fig. 1d (switch at +1).

In reactions containing single nucleotides, human Pol η primarily inserts dATP opposite both thymines of a dimer (see, for example, refs 1, 2, 9, 10). However, error rates have not been determined for dimer bypass reactions that allow competition among all four dNTPs and that require insertion opposite both damaged bases and continued polymerization. Therefore, we used a newly devel-

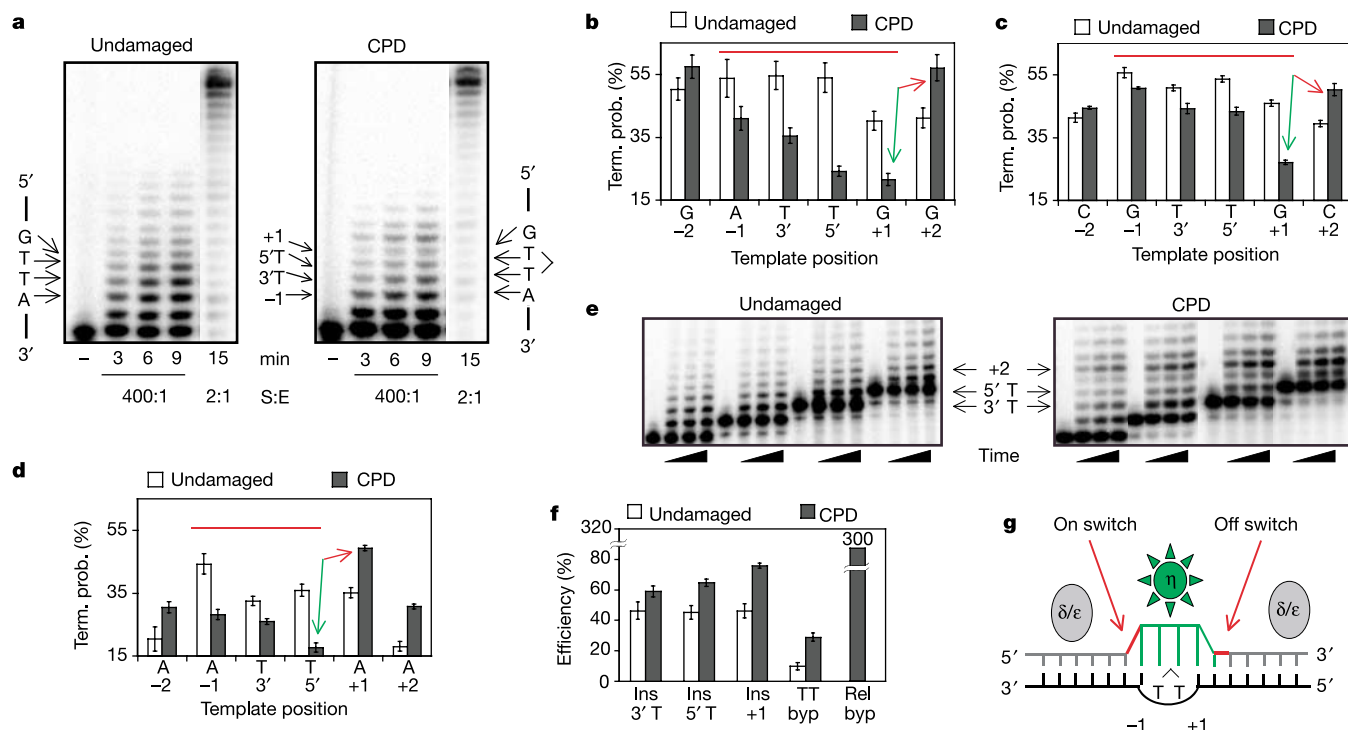


Figure 1 CPD bypass by Pol η . **a**, Primer extension reactions were done with the indicated substrate:enzyme (S:E) ratios and incubation times (min). **b–d**, Termination probabilities were calculated as described⁷ and represent the mean $\pm$ s.d. for 3–10 determinations per substrate. **e**, Primer extension reactions with primers of 26–29

nucleotides. **f**, Values for insertion (Ins) at three template positions, for bypass (byp) of undamaged thymine or the TT dimer, and for relative bypass (Rel byp) efficiency⁷.

g, Model of polymerase switching during CPD bypass.

oped system⁷ to measure error rates for a complete dimer bypass reaction. Templates were copied to generate primarily full-length DNA products (Figs 1a and 2a, right lanes) that were hybridized to gapped M13lac DNA molecules and used to obtain M13 plaques⁷. By using the template 5'-TTAG sequence context, base substitutions at the 3' T of the dimer (italicized) were easily scored with 60% expression in *Escherichia coli* as dark blue plaque revertants of the TAG amber codon in the *lacZ* α -complementation gene, which encodes a faint blue plaque phenotype (Supplementary Table S1). Base substitutions at the 3' and 5' T were identified by sequence analysis and plaque hybridization, respectively (Supplementary Table S1). Deletions and insertions were detected as colourless plaques and sequenced (Supplementary Table S1).

The results reveal several features that were reproducible in two separate experiments (Supplementary Table S1). Dimer bypass fidelity is low (Fig. 3a, b). For example, Pol η and Dpo4 incorporate one dGMP for every 26 or 10 dAMPs incorporated, respectively, opposite the 3' T of the dimer. For this same error, Pol η copies the 5' T of the dimer 4.4-fold more accurately than it copies the corresponding undamaged thymine (32×10^{-4} versus 140×10^{-4}). The difference with damaged DNA is similar to a 4.8-fold higher discrimination against dGTP mis-insertion opposite the 5' T of a dimer, as determined by steady-state kinetic analysis in reactions containing single nucleotides (see Table II in ref. 15). Both Pol η and Dpo4 also generate single-base deletion and addition errors at the location of the dimer at rates that are readily detectable (Fig. 3a, b). Formation of deletion errors is consistent with the ability of Dpo4 to accommodate an unpaired nucleotide in the active site¹⁴. Deletion and insertion errors by Pol η during dimer bypass could underlie some of the frame-shift mutagenesis associated with ultraviolet irradiation¹⁶.

For both human Pol η and Dpo4, bypass error rates differ at the 3' and 5' T of the dimer (Fig. 3). For Pol η , the error rate for T to C substitutions (T-dGTP mismatches) at the 5' T is 32×10^{-4} , a

value that is 12-fold lower than at the 3' T of the dimer. For Dpo4, the T to C substitution error rate at the 5' T is 67-fold lower than at the 3' T. Differences at these two positions with undamaged DNA are smaller, indicating that the biases in fidelity at the 3' and 5' T with the damaged template are largely due to the lesion rather than the local DNA sequence context.

We previously found that Pol η copies undamaged DNA with very low fidelity^{17,18} and suggested that, as compared with more accurate polymerases, Pol η fidelity may largely depend on free-energy differences between correct and incorrect base pairs determined by base-base hydrogen bonding. Notably, a DNA decamer containing a TT dimer⁸ has normal Watson-Crick hydrogen bonds at the 3' base pair, but its 5' base pair has lost one hydrogen bond and the other is longer than usual. In addition, the 3' T of a TT dimer forms a Watson-Crick base pair with ddATP at the active site of *S. solfataricus* Dpo4, whereas the 5' T of a dimer forms a Hoogsteen base pair with ddATP in its *syn* conformation³.

Thus, abnormal Watson-Crick and Hoogsteen base pairings are two possible solutions to correct templating by the 5' T, even though it is covalently linked to the preceding base. The higher fidelity of Dpo4 and Pol η at the 5' T of a dimer may result from lower stability of mismatches in an abnormal pairing configuration involving fewer or weaker base-base hydrogen bonds or altered base stacking. These and other explanations must eventually account for the fact that overall dimer bypass fidelity is low and that error rates differ depending on the polymerase, the mismatch and its position (Fig. 3). The dimer bypass error rate differences between Pol η and Dpo4 seen here may result from differences in conformational flexibility and/or active site geometry, as determined by specific amino acid differences (see also ref. 3).

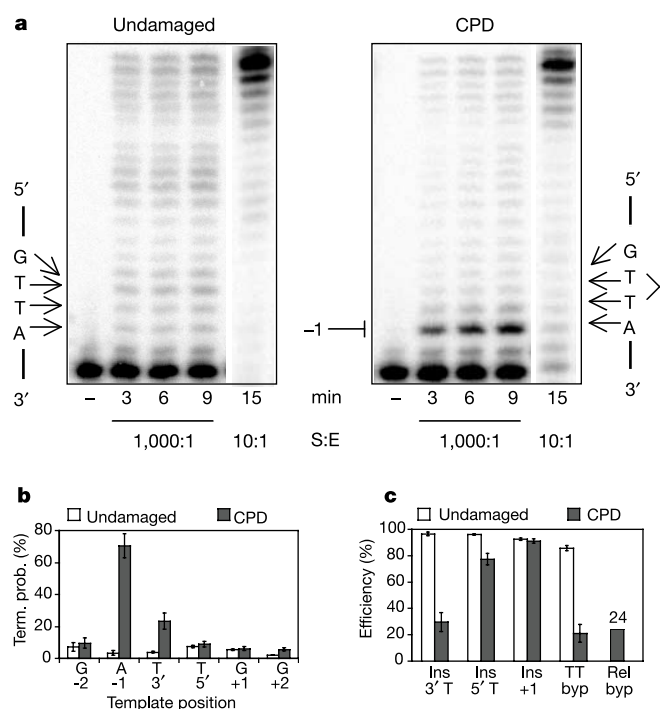


Figure 2 CPD bypass by *S. solfataricus* Dpo4. **a**, Primer extension reactions similar to those described in Fig. 1a were done with Dpo4. **b**, Termination probabilities were measured as in Fig. 1b. **c**, Insertion (Ins), bypass (byp) and relative efficiency (Rel byp) were measured as in Fig. 1f.

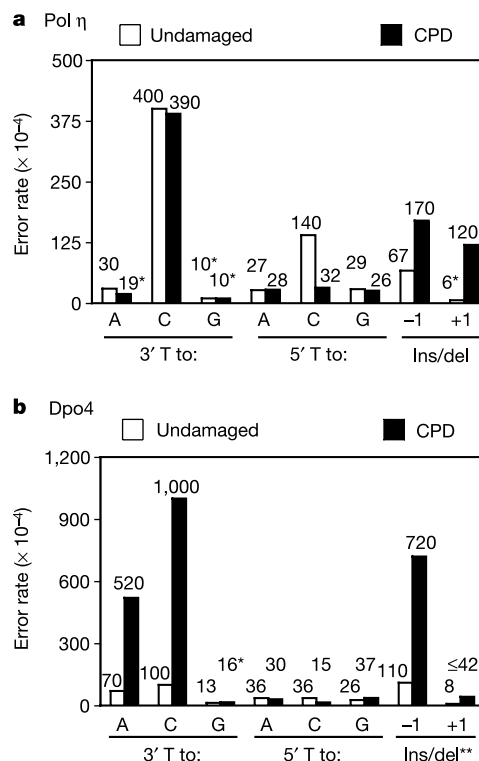


Figure 3 Error rates for undamaged thymine and TT dimer bypass. **a**, Human Pol η ; **b**, *S. solfataricus* Dpo4. Most error rates shown are the average of two experiments (Supplementary Table S1). In five cases (asterisk) where no mutant plaques containing a particular change were found in one of the two experiments, rates were calculated using the plaque total from both experiments. Frame-shift error rates for Dpo4 are from a single experiment (double asterisk).

Understanding the enzymatic origins of sunlight-induced mutagenesis in any human, with the xeroderma pigmentosum variant (XPV) gene or otherwise, is very complicated (see refs 16, 19), because sunlight induces many different lesions, including potentially mutagenic 6-4 photoproducts²⁰. In addition, humans encode at least 15 DNA template-dependent DNA polymerases^{12,21}. As noted for low-fidelity synthesis by Pol η with undamaged DNA¹⁷, the fidelity of dimer bypass involving Pol η *in vivo* may be higher than that seen *in vitro*, owing to proofreading by another protein (such as Pol δ or Pol ϵ), polymerase accessory proteins, protein modifications or DNA mismatch repair.

These possibilities and the likelihood that dimers will be encountered rarely in NER-proficient individuals with xeroderma pigmentosum can account for the suppression of sunlight-induced mutagenesis and skin cancer by Pol η despite its intrinsically low bypass fidelity. At the same time, it is notable that no other human DNA polymerase examined so far (α , β , δ , ζ , κ , ι and μ ; refs 9, 22–26) approaches the high dimer bypass efficiency of Pol η reported here. This implies that, when present in normal people and in NER-defective individuals with xeroderma pigmentosum, Pol η is the enzyme most likely to bypass a cyclobutane pyrimidine dimer (CPD). The inherently low fidelity of Pol η and the difference in fidelity at the 3' and 5' T of the dimer, combined with the fact that ultraviolet-induced mutations in cells and skin tumours from normal humans are characterized by higher mutagenesis at the 3' base than at the 5' base of dipyrimidine sequences^{4–6}, is consistent with the possibility that replication errors by Pol η may contribute to some of the mutagenesis associated with skin cancer (also see ref. 27). □

Methods

Bypass efficiency reactions

The primers, 5'-AATTTCTGTCAGGTCGACTCCAAAGG-3' (Figs 1a, b, and 2a–c), 5'-CACTTACTGTAGC-3' (Fig. 1c) and 5'-AGCTATGACCATGATTACGAATTGC-3' (Fig. 1d), were labelled with ³²P at the 5' end using T4 polynucleotide kinase and annealed to a 1.5 molar excess of either an undamaged or a CPD-containing oligonucleotide (CPD position is italicized) with the following sequence: 5'-CCAGCTCGGTACCGGGTT AGCCTTTGAGTGCAGCTGCAGAAATT-3' (Figs 1a, b, e, and 2a–c), 5'-GGCTTGTCACATCGCGTT GCGCTACAGTAAGTG-3' (Fig. 1c) or 5'-AGCTACCATGCCTGCACGAATT AAGCAATTCGTAATCATGGTCATAGCT-3' (Fig. 1d). Primers in Fig. 1e are similar to those in Fig. 1a, b, but are 1–4 nucleotides longer at the 3' end. We synthesized CPD-containing oligonucleotides as described²⁸.

Pol η and Dpo4 were purified as described^{9,11}. Reactions with Pol η (30 μ M, 37 °C) contained 40 mM Tris-HCl (pH 8.0), 10 mM MgCl₂, 60 mM KCl, 100 μ M each of the four dNTPs, 10 mM DTT, 2.5% glycerol, 250 μ g ml⁻¹ bovine serum albumin (BSA), 130 nM primer template and 0.33 nM enzyme. Reactions with Dpo4 (30 μ M, 70 °C) contained 40 mM Tris-HCl pH 9.0 (at 25 °C), 5 mM MgCl₂, 100 μ M each of the four dNTPs, 10 mM DTT, 2.5% glycerol, 250 μ g ml⁻¹ BSA, 130 nM primer template and 0.13 nM enzyme. Samples (6 μ l) were removed at the indicated times, added to an equal volume of formamide stop solution (95% deionized formamide, 25 mM EDTA, 0.01% bromophenol blue and 0.01% xylene cyanol), heated to 95 °C for 3 min, resolved by 12% denaturing PAGE and quantified by phosphorimager. We calculated termination probability, insertion efficiency, extension efficiency and bypass efficiency as described⁷.

Bypass fidelity reactions

Reactions were done as described above with 65 nM Pol η or 13 nM Dpo4 and unlabelled primer templates in the presence of [³²P- α]dCTP. Reaction products were processed and annealed to gapped M13lac DNA as described⁷. M13 DNA molecules were introduced into *E. coli* and mutant frequencies and error rates for the 3' T and for frame-shift errors were calculated as described¹⁸. Error rates at the 5' T were determined by plaque hybridization of nylon membrane lifts from plates containing 200–500 plaques. Membranes were hybridized to the following ³²P end-labelled oligonucleotides: CPD-AG (T $\rightarrow$ C), 5'-AAGGCTAGCCCGTA-3'; CPD-AC (T $\rightarrow$ G), 5'-AAGGCTACCCCGTA-3'; CPD-AT (T $\rightarrow$ A), 5'-AAGGCTATCCCGTA-3'. Membranes were washed for 5 min each at 25, 37 and 48 °C with 6 $\times$ SSC buffer and visualized by phosphorimaging. We calculated error rates as (N/P_i)/E, where N is the number of mutant plaques observed, P_i is the total plaques screened, and E is 60% (the efficiency of expressing polymerase errors on transfection of *E. coli* cells⁷).

Received 5 November 2003; accepted 15 January 2004; doi:10.1038/nature02352.

- Masutani, C. *et al.* The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η . *Nature* **399**, 700–704 (1999).
- Johnson, R. E., Kondratieff, C. M., Prakash, S. & Prakash, L. hRAD30 mutations in the variant form of xeroderma pigmentosum. *Science* **285**, 263–265 (1999).
- Ling, H., Boudsocq, F., Plosky, B. S., Woodgate, R. & Yang, W. Replication of a *cis-syn* thymine dimer at atomic resolution. *Nature* **424**, 1083–1087 (2003).
- Armstrong, J. D. & Kunz, B. A. Site and strand specificity of UVB mutagenesis in the *SUP4-o* gene of yeast. *Proc. Natl Acad. Sci. USA* **87**, 9005–9009 (1990).
- Dumaz, N., Stary, A., Soussi, T., Daya-Grosjean, L. & Sarasin, A. Can we predict solar ultraviolet radiation as the causal event in human tumours by analyzing the mutation spectra of the p53 gene? *Mutat. Res.* **307**, 375–386 (1994).
- Hsia, H. C., Lebkowski, J. S., Leong, P.-M., Calos, M. P. & Miller, J. H. Comparison of ultraviolet irradiation-induced mutagenesis of the *lacI* gene in *Escherichia coli* and in human 293 cells. *J. Mol. Biol.* **205**, 103–113 (1989).
- Kokoska, R. J., McCulloch, S. D. & Kunkel, T. A. The efficiency and specificity of apurinic/apyrimidinic site bypass by human DNA polymerase η and *Sulfolobus solfataricus* Dpo4. *J. Biol. Chem.* **278**, 50537–50545 (2003).
- Masutani, C., Kusumoto, R., Iwai, S. & Hanaoka, F. Mechanisms of accurate translesion synthesis by human DNA polymerase η . *EMBO J.* **19**, 3100–3109 (2000).
- Park, H. *et al.* Crystal structure of a DNA decamer containing a *cis-syn* thymine dimer. *Proc. Natl Acad. Sci. USA* **99**, 15965–15970 (2002).
- Washington, M. T., Prakash, L. & Prakash, S. Mechanism of nucleotide incorporation opposite a thymine–thymine dimer by yeast DNA polymerases η . *Proc. Natl Acad. Sci. USA* **100**, 12093–12098 (2003).
- Boudsocq, F., Iwai, S., Hanaoka, F. & Woodgate, R. *Sulfolobus solfataricus* P2 DNA polymerase IV (Dpo4): an archaeal DinB-like DNA polymerase with lesion-bypass properties akin to eukaryotic pol η . *Nucleic Acids Res.* **29**, 4607–4616 (2001).
- Shcherbakova, P. V., Bebenek, K. & Kunkel, T. A. Functions of eukaryotic DNA polymerases. *Sci. Aging Knowledge Environ.* [online], 26 February 2003 (sageke.sciencemag.org/cgi/content/full/sageke;2003/8/rev3).
- Trincão, J. *et al.* Structure of the catalytic core of *S. cerevisiae* DNA polymerase η : implications for translesion DNA synthesis. *Mol. Cell* **8**, 417–426 (2001).
- Ling, H., Boudsocq, F., Woodgate, R. & Yang, W. Crystal structure of a Y-family DNA polymerase in action: a mechanism for error-prone and lesion-bypass replication. *Cell* **107**, 91–102 (2001).
- Johnson, R. E., Washington, M. T., Prakash, S. & Prakash, L. Fidelity of human DNA polymerase η . *J. Biol. Chem.* **275**, 7447–7450 (2000).
- Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* 544–555 (ASM, Washington DC, 1995).
- Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F. & Kunkel, T. A. Low fidelity DNA synthesis by human DNA polymerase- η . *Nature* **404**, 1011–1013 (2000).
- Matsuda, T. *et al.* Error rate and specificity of human and murine DNA polymerase η . *J. Mol. Biol.* **312**, 335–346 (2001).
- Prakash, S. & Prakash, L. Translesion DNA synthesis in eukaryotes: a one- or two-polymerase affair. *Genes Dev.* **16**, 1872–1883 (2002).
- Bressan, A. & Fuchs, R. P. Lesion bypass in yeast cells: Pol η participates in a multi-DNA polymerase process. *EMBO J.* **21**, 3881–3887 (2002).
- Marini, F., Kim, N., Schuffert, A. & Wood, R. D. POLN, a nuclear PolA family DNA polymerase homologous to the DNA cross-link sensitivity protein Mus308. *J. Biol. Chem.* **278**, 32014–32019 (2003).
- O'Day, C. L., Burgers, P. M. J. & Taylor, J. PCNA-induced DNA synthesis past *cis-syn* and *trans-syn-I* thymine dimers by calf thymus DNA polymerase δ *in vitro*. *Nucleic Acids Res.* **20**, 5403–5406 (1992).
- Tissier, A. *et al.* Misinsertion and bypass of thymine–thymine dimers by human DNA polymerase ι . *EMBO J.* **19**, 5259–5266 (2000).
- Nelson, J. R., Lawrence, C. W. & Hinkle, D. C. Thymine–thymine dimer bypass by yeast DNA polymerase ζ . *Science* **272**, 1646–1649 (1996).
- Zhang, Y. *et al.* Lesion bypass activities of human DNA polymerase μ . *J. Biol. Chem.* **277**, 44582–44587 (2002).
- Ohashi, E. *et al.* Error-prone bypass of certain DNA lesions by the human DNA polymerase κ . *Genes Dev.* **14**, 1589–1594 (2000).
- Zhang, H. & Siede, W. UV-induced T $\rightarrow$ C transition at a TT photoproduct site is dependent on *Saccharomyces cerevisiae* polymerase η *in vivo*. *Nucleic Acids Res.* **30**, 1262–1267 (2002).
- Murata, T., Iwai, S. & Ohtsuka, E. Synthesis and characterization of a substrate for T4 endonuclease V containing a phosphorodithioate linkage at the thymine dimer site. *Nucleic Acids Res.* **18**, 7279–7286 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Y. Pavlov and K. Bebenek for discussions and comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.A.K. (kunkel@niehs.nih.gov).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Grad school confidential

In the classic underground comic strip *Art School Confidential*, by Daniel Clowes (creator of the *Ghost World* comic book and screenplay), one art-school instructor levels with her students, saying that only 10% of them will ever have successful careers in their field of choice. Re-reading this strip recently, I was struck by the resemblance of this message to what I would like to see science advisers at graduate schools say to their students and postdocs early on about the odds that they will land an academic job, post-PhD (see *Naturejobs* 3; 22 August 2002).

But I was also struck by the way in which the comic-strip art students reacted to that discouraging statistic. All of them had 'thought bubbles' floating over their heads filled with "It's going to be me!". One could read such sentiments — both for art students and science graduates — in two ways. The cynic would say: "What a terribly misguided, ultimately doomed bunch." The optimist would say: "It's great to see people pursue their dreams."

I put myself firmly in the optimist's camp. Science graduate students need to believe that they have a chance of success. Otherwise, why would they endure the daily grind of graduate school and the uncertainty of ever landing a permanent position?

So when the comic strip turns into a major motion picture later this year, science graduate students who go to watch it will imagine that its lead actor, John Malkovich, is wearing a lab coat rather than a smock. And they will probably nod in solidarity with their humanities viewers when the strip's filmic equivalent tells the student characters about their long odds. And, in all probability, art students and science students alike will vow to beat them.

Paul Smaglik
Naturejobs editor



Contents

POSTDOCS & STUDENTS

Life on a shoestring budget p102

CAREER VIEW

Nuts & Bolts
Reference points
Graduate Journal
Working for balance
Movers
Sun Kwok p104

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Stipend survival

Graduate-student pay levels mean tight budgets and inventive cost-cutting, but is the five-year pay freeze worth it? Kendall Powell calculates the bottom line.

Matt McNatt drives what he describes as a “third-generation, family hand-me-down” 1989 Chevrolet Caprice with 212,000 miles on the clock. It’s unreliable, so most days the third-year graduate student in cell biology bikes to his lab at the University of Colorado, Boulder. McNatt brings in US\$1,791 a month and says one-third of that used to go on rent before he found more people to share housing costs with.

All over the world, graduate students stretch their take-home pay to cover daily living expenses. To make ends meet, students forgo or share cars, limit shopping to sales and take advantage of free campus activities. The low income leaves little room for savings or an extravagant lifestyle. While college friends may go on to high-paying ‘real’ jobs, graduate students face five years or more of living hand to mouth.

Not surprisingly, graduate student spending closely follows psychologist Abraham Maslow’s pyramid of needs: basic needs come first, followed by social needs — if there’s money left over. Most PhD candidates, however, find that the top two tiers of the pyramid, ego needs and self-actualization, remain out of reach.

LIVING LIKE KINGS ON THE CHEAP

Shelter is on the foundation layer of Maslow’s pyramid, and for most graduate students that’s what eats up the bulk of their pay. Buying a house is generally an unattainable goal, much less a source of additional income. But when Brook Brouha entered

Tale of two stipends

In 2003, the US National Science Foundation (NSF) raised the graduate stipend for its training grant programme to \$25,000, making it competitive with the highest-paying departments in the country. Meanwhile, the US National Institutes of Health (NIH) graduate fellowship stipend for 2003 was \$19,968. Jade Zee, a neuroscience PhD candidate at the University of Oregon in Eugene, says an awkward situation arose in her department, with graduate students in the same lab receiving different pay because of the different scope of their projects.

The NIH’s research training officer Walter Schaffer says his programme has set a goal of bringing the graduate stipend up to \$25,000 by 2006. Schaffer admits that the programme lags behind other federal graduate fellowships. But the NIH funded 69% of all federal graduate awards in 2003, shouldering a much heavier burden than all other agencies.

The NSF, which funded 15% of federal awards last year, has taken a more aggressive stance. “We are trying to make going to graduate school competitive with other jobs,” says Lenore Clesceri, acting programme director for NSF training grants. The 2005 stipend has been set at \$30,000.

K.P.



M. JADE ZEE

his MD/PhD programme at the University of Pennsylvania in Philadelphia about eight years ago, he shrewdly put \$500 down on a house in a “slightly sketchy” area near the university. As a first-time buyer, he soon had a two-bedroom house with a \$600 per month mortgage and rented out one bedroom for \$425. Recently, Brouha moved out, took advantage of the lower interest rates to refinance the house, brought the mortgage down to \$200 per month, then rented it to new tenants for \$1,300 a month.

“The house made me rich in terms of grad student life,” says Brouha, now in his third year of medical school at the University of Pennsylvania. He also says the big financial risks of buying a house and investing in the stock market helped him take scientific risks in the lab in his stride. “Money never meant anything more than a concept — you are as rich as you feel,” says Brouha. “Ninety per cent of success is just walking in the door each day.”



M. MCNATT/B. BROUHA

Principles of parsimony, from left: Jade Zee eats at recruitment dinners to save the food budget for her four-legged friends; Matt McNatt drives an elderly car that doesn't always get there; Brook Brouha funds skiing trips through a wise housing investment.

Others look for ways to save on this biggest of expenses. Jade Zee, a sixth-year neuroscience student at the University of Oregon in Eugene, makes about \$19,000 a year, but saves money on rent by living in subsidized graduate student housing (see *Nature* **421**, 766–767; 2003). Frederic Laporte, a second-year student at McGill University in Montreal, Canada, lived in an ultra-low-rent district and saved enough to purchase a condominium this year.

His labmate Catherine Au, a third-year student, says that the province's student-friendly stance and the city's low cost of living drew her to the university. "My choice in McGill was that it is a first-class university and there was no financial burden." The low cost of living in Montreal means that their stipend of about Can\$14,500 (US\$10,800) goes pretty far.

Transport comes a close second in the list of expenses, especially in places without a good public system. Andrea Sweigert jumped on an ad in her department for a free old car. She and another graduate student at Duke University in Durham, North Carolina, share it, using it to "go to Chapel Hill or go out on weekends", she says. "It floats between our houses with a bike rack on the back and it works really well."

Besides housing and cars, students find other ways to pinch pennies, such as shopping in the sales and being creative with leftovers. And those pennies go into the social calendar fund. Brouha has shown that, using a bit of cleverness and taking some risks, it's possible for a postgrad to climb Maslow's pyramid at least as high as ego gratification. He used his rental income to make trips to a friend's house in Utah and go skiing nearby. Laporte manages about 20 rounds of golf a summer. The key, he says, is to choose cheap courses and tee times and ask for new equipment as

Christmas presents. Zee says she takes advantage of free activities such as local hiking and department recruiting dinners. She spends her savings on her two biggest extravagances — Luna the black labrador and Jasmine the tuxedo cat.

WORTH VERSUS WEALTH

But not everyone gets by so swimmingly. In Sweigert's biology department, the \$16,000 teaching assistantships cover nine months' salary, and only about half the labs pay students during the summer. Sweigert says students resort to extra teaching, mentoring jobs, or even being a dorm adviser to make ends meet. "They are spending 20 hours a week on something else, not working on their projects," she says.

Stipends in ecology and evolutionary biology have typically been lower than in the biomedical sciences. It's not uncommon for students to make up the difference with unsanctioned, non-campus work such as babysitting, yoga instruction or bartending. The discrepancy between biomedicine and the rest could be set to do an about-face, however, as a major US funding agency for non-biomedical science training recently set a markedly higher stipend level (see 'A tale of two stipends').

Even those doing reasonably well wouldn't turn down a pay rise. Laporte acknowledges that, although he and many others see the system as somewhat exploitative, it would take "a worldwide strike" to change it. Many say that, compared to other professions, it's hard to tell when training ends and cheap labour begins. "For the high qualification — after almost a decade at university — salaries are miserable, even as a postdoc," says Christian Haering, a postdoc at the Institute of Molecular Pathology (IMP) in Vienna, Austria, where he also completed his PhD. Even so, it's difficult to find any 'starving students' in the sciences. In fact, the IMP, funded by a private company, pays postgraduates a salary (with benefits) that is about €2,000 (US\$2,542) more than stipends at Austrian universities.

In China, Rong Xu, a fourth-year genetics student at Fudan University in Shanghai, reports that his monthly stipend of 1,450 yuan (US\$175) puts him just below the average city worker's income of 2,000 yuan. Stipends at the Fudan–Yale Biomedical Research Center include supplements for housing and for lunch and dinner on days spent in lab.

TOP OF THE HEAP

But intellectual freedom and originality of research rank high among most postgrads' explanations for why getting a doctorate is worth the low pay. Xu's colleague, Congwu Chi, knows that you don't need to get to the top of the mountain to reach the top of the pyramid. He realizes he could make a little more money working at a company, but says he can learn more in an academic lab. "In the company, I might only be responsible for one part of the work and just repeat this part every day."

Kendall Powell is a freelance science writer based in Broomfield, Colorado.

NUTS & BOLTS

GRADUATE JOURNAL

Working for balance

We all know those dark days of graduate work. A three-week experiment has failed. You dropped your perfect gel. The questions after your presentation made you feel as if you had the brains of a hamster. You think of leaving and opening a fast-food franchise.

I have learned the importance of including something else in my life to keep me from disintegrating after a dark day of science. This something else must wipe all thoughts of the day of disaster from my mind. Justifying the time spent on extra-curricular pursuits to one's supervisor can be daunting, but some labs incorporate these pursuits into everyday life (see *Nature* **427**, 268–269; 2004).

Rowing indirectly brought me to my graduate life, and it still propels me through days of data despair. If I fail to focus on balancing the boat, I am certain to become a soggy scientist. Biological chemistry graduate student Josh Finkelstein hunts for harmony with his band of rock'n'roll chemists, and physics postdoc Etienne Boaknin lost and found his balance on the judo mat. The time and energy invested in our hobbies clears our minds, gives us something else to have nightmares about, and provides another subject for conversation in job interviews. ■

Sidney Omelon is a PhD student in bone biomaterials at the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada.

Reference points

Lining up references goes hand-in-hand with making a résumé or CV. In the final stage of the selection process, recommendations can make or break a candidacy. To make your references as effective as they can be, follow these three steps.

First, your referees should know your background, have seen your work close up and hold you in high regard. Employers want their good impressions reinforced. Former advisers who will enthuse about your qualifications are ideal. Asking someone to serve as a referee may seem like an obvious first step, but be aware that every day some unsuspecting graduate adviser or laboratory supervisor gets called up out of the blue to comment on the performance of a past advisee or employee. Approach your carefully



With Deb Koen
Careers consultant

selected targets to ask if they would be willing to serve as referees and if they would be comfortable in recommending you strongly.

Second: your responsibility doesn't end with the selection and request. When you've got your list, create a reference sheet to distribute to potential employers upon request, usually after an interview and after finalists have emerged. Most employers prefer to check references on the telephone, although some, particularly in academia, may ask for letters. These

should be tailored to the specific openings sought. Don't assume that your referees remember everything about you — they need guidance. Extend the courtesy of a call or an e-mail with details about the position. Remind them of the skills and accomplishments that are most relevant to your current focus.

Finally, after your referees have taken the time to speak or write on your behalf, let them know how your career unfolds. Express your appreciation and provide updates. Try not to overuse them, and only provide your reference list when requested. Be sure to share the good news when you get a job, and keep the door open for future opportunities. After all, you never know when you might need a glowing reference again. ■

Deb Koen is vice-president of Career Development Services and a columnist for *The Wall Street Journal's CareerJournal.com*.

MOVERS

Sun Kwok, director, Academia Sinica Institute of Astronomy and Astrophysics, Taiwan



One book changed Sun Kwok's career. While an undergraduate at McMaster University in Canada, the Hong Kong-born engineering major picked up a copy of Fred Hoyle's *Frontiers of Astronomy*. Until then, he had associated astronomy with sophisticated star-gazing, but as he turned the pages he realized that modern physics had changed the discipline. By the time he put the book down, Kwok knew he wanted to change majors.

When he began his PhD at the University of Minnesota in Minneapolis, Kwok was able to take advantage of the

expanding physical tool kit Hoyle's book proposed. "In Minnesota we were fortunate because they had the first telescope with mid-infrared detectors," he says. That helped Kwok in his early work on stellar winds from red giants.

But his curiosity expanded beyond these large, mature stars. So in 1976, the self-described "red-giant guy" attended an International Astronomical Union symposium on planetary nebulae. At the conference he learned that the conventional theories of sudden ejection of the nebulae had lots of problems. So his knowledge of red-giant winds led him to propose in 1978 that planetary nebulae form when stellar winds plough pre-existing planetary gas out of the way, piling it up like a snowbank.

As the physical tools of astronomy improved, Kwok's theoretical model was validated with infrared, ultraviolet and X-ray observations. More recently, technology helped attract Kwok to

Taiwan. His relationship with the country extends back to 1991, when he helped Frank Shu to develop a ten-year plan to build up its astronomy.

Now at the helm of the Academia Sinica Institute of Astronomy and Astrophysics, Kwok is in a position to establish Taiwan's role in two international astronomy projects. The Submillimeter Array — eight antennas on Mauna Kea in Hawaii that can image molecular emissions from star-forming regions — is already under way. And Kwok hopes to get Taiwan involved in an even more ambitious project — the Atacama Large Millimeter Array, an international project currently led by US, Japanese and European astronomers.

All of these developments mean more tools for the next generation of astronomers to use. "We now have all these capabilities to look at the Universe in a new light," Kwok says. "It's still very much a young and developing field." ■

CV

2000–02: Killam Fellow, Canada Council for the Arts

1983–2003: Professor at the University of Calgary, Canada

1978–83: Research associate at the Herzberg Institute of Astrophysics, Ottawa, Canada

1974: PhD, University of Minnesota, Minneapolis